A

DOCKET NO. : IBIS-0012 **PATENT**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In Re Application of:

David J. Ecker, Ranga Sampath, Richard Griffey and John McNeil

Serial No.: Not Yet Assigned **Group Art Unit:** Not Yet Assigned

Filing Date: Herewith **Examiner:** Not Yet Assigned

For: IDENTIFICATION OF MOLECULAR INTERACTION SITES IN RNA FOR NOVEL DRUG DISCOVERY

EXPRESS MAIL LABEL NO: EL066387917US
DATE OF DEPOSIT: May 12, 1999

Box ☒ Patent Application

☐ Provisional ☐ Design ☐ Sequence

Assistant Commissioner for Patents
Washington DC 20231

Sir:

PATENT APPLICATION TRANSMITTAL LETTER

Transmitted herewith for filing, please find

☒ A Utility Patent Application under 37 C.F.R. 1.53(b).

It is a continuing application, as follows:

☐ continuation ☐ divisional ☒ continuation-in-part of prior application number 09 /076,440, filed May 12, 1998.

☐ A Provisional Patent Application under 37 C.F.R. 1.53(c).

☐ A Design Patent Application (submitted in duplicate).

Including the following:

☐ Provisional Application Cover Sheet.

☒ New or Revised Specification, including pages 1 to 49 containing:

☒ Specification

☒ Claims

☒ Abstract

☐ Substitute Specification, including Claims and Abstract.

☐ The present application is a continuation application of Application No. _____ filed _____. The present application includes the Specification of the parent application which has been revised in accordance with the amendments filed in the parent application. Since none of those amendments incorporate new matter into the parent application, the present revised Specification also does not include new matter.

☐ The present application is a continuation application of Application No. _____ filed _____, which in turn is a continuation-in-part of Application No. _____ filed _____. The present application includes the Specification of the parent application which has been revised in accordance with the amendments filed in the parent application. Although the amendments in the parent C-I-P application may have incorporated new matter, since those are the only revisions included in the present application, the present application includes no new matter in relation to the parent application.

☐ A copy of earlier application Serial No. _____ Filed _____, including Specification, Claims and Abstract (pages 1 - @@), to which no new matter has been added TOGETHER WITH a copy of the executed oath or declaration for such earlier application and all drawings and appendices. Such earlier application is hereby incorporated into the present application by reference.

☐ Please enter the following amendment to the Specification under the Cross-Reference to Related Applications section (or create such a section) : "This Application: ☐ is a continuation of ☐ is a divisional of ☐ claims benefit of U.S. provisional Application Serial No. _____filed _____

_____

_____

_____.

☐ Signed Statement attached deleting inventor(s) named in the prior application.

☐ A Preliminary Amendment.

☒ 100 Sheets of ☒ Formal ☐ Informal Drawings.

☐ Petition to Accept Photographic Drawings.

☐ Petition Fee

☒ An ☐ Executed ☒ Unexecuted Declaration or Oath and Power of Attorney.

☐ An Associate Power of Attorney.

☐ An ☐ Executed ☐ Copy of Executed Assignment of the Invention to _____

_____.

☐ A Recordation Form Cover Sheet.

☐ Recordation Fee - $40.00.

☒ The prior application is assigned of record to ISIS Pharmaceuticals, Inc.

☐ Priority is claimed under 35 U.S.C. § 119 of Patent Application No. _____ filed _____ in _____ (country).

☐ A Certified Copy of each of the above applications for which priority is claimed:

☐ is enclosed.

☐ has been filed in prior application Serial No. _____ filed _____.

☒ An ☒ Executed or ☐ Copy of Executed Earlier Statement Claiming Small Entity Status under 37 C.F.R. 1.9 and 1.27

☒ is enclosed.

☐ has been filed in prior application Serial No. _____ filed _____,

said status is still proper and desired in present case.

☐ Diskette Containing DNA/Amino Acid Sequence Information.

☐ Statement to Support Submission of DNA/Amino Acid Sequence Information.

☐ The computer readable form in this application _____, is identical with that filed in Application Serial Number _____, filed _____. In accordance with 37 CFR 1.821(e), please use the ☐ first-filed, ☐ last-filed or ☐ only computer readable form filed in that application as the computer readable form for the instant application. It is understood that the Patent and Trademark Office will make the necessary change in application number and filing date for the computer readable form that will be used for the instant application. A paper copy of the Sequence Listing is ☐ included in the originally-filed specification of the instant application, ☐ included in a separately filed preliminary amendment for incorporation into the specification.

☐ Information Disclosure Statement.

- ☐ Attached Form 1449.
- ☐ Copies of each of the references listed on the attached Form PTO-1449 are enclosed herewith.

☐ A copy of Petition for Extension of Time as filed in the prior case.

☐ Appended Material as follows: _____.

☒ Return Receipt Postcard (should be specifically itemized).

☐ Other as follows: _____

_____

_____.

FEE CALCULATION:

☐ Cancel in this application original claims _____ of the prior application before calculating the filing fee. (At least one original independent claim must be retained for filing purposes.)

			SMALL ENTITY		NOT SMALL ENTITY	
			RATE	FEE	RATE	FEE
PROVISIONAL APPLICATION			$75.00	$	$150.00	$
DESIGN APPLICATION			$155.00	$	$310.00	$
UTILITY APPLICATIONS BASE FEE			$380.00	$380.00	$760.00	$
UTILITY APPLICATION; ALL CLAIMS CALCULATED AFTER ENTRY OF ALL AMENDMENTS						
	No. Filed	No. Extra				
TOTAL CLAIMS	34- 20 =	14	$9 each	$126	$18 each	$
INDEP. CLAIMS	6- 3 =	3	$39 each	$117	$78 each	$
FIRST PRESENTATION OF MULTIPLE DEPENDENT CLAIM			$130	$0	$260	$
ADDITIONAL FILING FEE				$0		$
TOTAL FILING FEE DUE				$623.00		$

☒ A Check is enclosed in the amount of $ 623.00 .

☒ The Commissioner is authorized to charge payment of the following fees and to refund any overpayment associated with this communication or during the pendency of this application to deposit account 23-3050. This sheet is provided in duplicate.

☐ The foregoing amount due.

☒ Any additional filing fees required, including fees for the presentation of extra claims under 37 C.F.R. 1.16.

☒ Any additional patent application processing fees under 37 C.F.R. 1.17 or 1.20(d).

☐ The issue fee set in 37 C.F.R. 1.18 at the mailing of the Notice of Allowance.

☒ The Commissioner is hereby requested to grant an extension of time for the appropriate length of time, should one be necessary, in connection with this filing or any future filing submitted to the U.S. Patent and Trademark Office in the above-

identified application during the pendency of this application. The Commissioner is further authorized to charge any fees related to any such extension of time to deposit account 23-3050. This sheet is provided in duplicate.

SHOULD ANY DEFICIENCIES APPEAR with respect to this application, including deficiencies in payment of fees, missing parts of the application or otherwise, the United States Patent and Trademark Office is respectfully requested to promptly notify the undersigned.

Date: 12 May 1999

Paul K. Legaard
Registration No. 38,534

Woodcock Washburn Kurtz
Mackiewicz & Norris LLP
One Liberty Place - 46th Floor
Philadelphia PA 19103
Telephone: (215) 568-3100
Facsimile: (215) 568-3439

Applicant or Patentee: David J. Ecker, Ranga Sampath, Richard Griffey, John McNeil

Serial No.: Not yet Known **Attorney's Docket No.: IBIS-0012**

Date Filed: Herewith

For: IDENTIFICATION OF MOLECULAR INTERACTION SITES IN RNA FOR NOVEL DRUG DISCOVERY

VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY STATUS (37 CFR 1.9(f) and 1.27(c) - SMALL BUSINESS CONCERN

I hereby declare that I am:

( ) the owner of the small business concern identified below:

(X) an official empowered to act on behalf of the concern identified below:

NAME OF CONCERN: ISIS Pharmaceuticals, Inc.
ADDRESS OF CONCERN: 2292 Faraday Avenue
Carlsbad, CA 92008

I hereby declare that the above-identified small business concern qualifies as a small business concern as defined in 13 CFR 121.12, and reproduced in 37 CFR 1.9(d), for purposes of paying reduced fees under section 41(a) and (b) of Title 35, United States Code, in that: (1) the number of employees of the concern, including those of its affiliates, does not exceed 500 persons; and (2) the concern has not assigned, granted, conveyed, or licensed, and is under no obligation under contract or law to assign, grant, convey, or license, any rights in the invention to any person who could not be classified as an independent inventor if that person had made the invention, or to any concern which would not qualify as a small business concern or a nonprofit organization under this section. For purposes of this statement, (1) the number of employees of the business concern is the average over the previous fiscal year of the concern of the persons employed on a full-time, part-time or temporary basis during each of the pay periods of the fiscal year, and (2) concerns are affiliates of each other when either, directly or indirectly, one concern controls or has the power to control the other, or a third party or parties controls or has the power to control both.

I hereby declare that rights under contract or law have been conveyed to and remain with the small business concern identified above with regard to the invention entitled **IDENTIFICATION OF MOLECULAR INTERACTION SITES IN RNA FOR NOVEL DRUG DISCOVERY**
by inventor(s)
described in

(XX) specification filed herewith.

( ) application serial no. _____, filed _____.

( ) patent no. _____, issued _____.

If the rights held by the above-identified small business concern are not exclusive, each individual, concern or organization having rights in the invention is listed below* and no rights to the invention are held by any person, other than the inventor, who would not qualify as an independent inventor under 37 CFR 1.9(c) if that person made the invention, or by any concern which would not qualify as a small business concern under 37 CFR 1.9(d), or a nonprofit organization under 37 CFR 1.9(e).

***NOTE: Separate verified statements are required for each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)**

FULL NAME:

ADDRESS:

( ) INDIVIDUAL ( ) SMALL BUSINESS CONCERN ( ) NONPROFIT ORGANIZATION

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b))

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

NAME OF PERSON SIGNING B. Lynne Parshall, Esq.

TITLE OF PERSON SIGNING Executive Vice President and Chief Financial Officer

ADDRESS OF PERSON SIGNING 2292 Faraday Avenue, Carlsbad, CA 92008

SIGNATURE

May 12, 1999

DATE

IDENTIFICATION OF MOLECULAR INTERACTION SITES IN RNA FOR NOVEL DRUG DISCOVERY

CROSS REFERENCE TO RELATED APPLICATIONS

The present application is a continuation-in-part of U.S. Serial No. 09/076,440 filed May 12, 1998, which claims priority to provisional U.S. Serial No. 60/085,092 filed May 12, 1998, each of which is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

The present invention is directed to methods of identifying regions of nucleic acids, especially RNA, in prokaryotes and eukaryotes that can serve as molecular interaction sites. Therapeutics and structural databases are also comprehended by the present invention.

BACKGROUND OF THE INVENTION

Recent advances in genomics, molecular biology, and structural biology have highlighted how RNA molecules participate in or control many of the events required to express proteins in cells. Rather than function as simple intermediaries, RNA molecules actively regulate their own transcription from DNA, splice and edit mRNA molecules and tRNA molecules, synthesize peptide bonds in the ribosome, catalyze the migration of nascent proteins to the cell membrane, and provide fine control over the rate of translation of messages. RNA molecules can adopt a variety of unique structural motifs, which provide the framework required to perform these functions.

"Small" molecule therapeutics, which bind specifically to structured RNA molecules, are organic chemical molecules which are not polymers. "Small" molecule therapeutics include the most powerful naturally-occurring antibiotics. For example, the aminoglycoside and macrolide antibiotics are "small" molecules that bind to defined regions in ribosomal RNA (rRNA) structures and work, it is believed, by blocking conformational changes in the RNA required for protein synthesis. Changes in the conformation of RNA molecules have been shown to regulate rates of transcription and translation of mRNA molecules.

An additional opportunity in targeting RNA for drug discovery is that cells frequently create different mRNA molecules in different tissues that can be translated into identical proteins. Processes such as alternative splicing and alternative polyadenylation can create transcripts that are unique or enriched in particular tissues. This provides the opportunity to design drugs that bind to the region of RNA unique in a desired tissue, including tumors, and not affect protein expression in other tissues, or affect protein expression to a lesser extent, providing an additional level of drug specificity generally not achieved by therapeutic targeting of proteins.

RNA molecules or groups of related RNA molecules are believed by Applicants to have regulatory regions that are used by the cell to control synthesis of proteins. The cell is believed to exercise control over both the timing and the amount of protein that is synthesized by direct, specific interactions with mRNA. This notion is inconsistent with the impression obtained by reading the scientific literature on gene regulation, which is highly focused on transcription. The process of RNA maturation, transport, intracellular localization and translation are rich in RNA recognition sites that provide good opportunities for drug binding. Applicants' invention is directed to finding these regions for RNA molecules in the human genome as well as in other animal genomes and prokaryotic genomes.

Accordingly, it is a principal object of the invention to identify molecular interaction sites in nucleic acids, especially RNA. A further object of the invention is to identify secondary structural elements in RNA which are highly likely to give rise to significant therapeutic, regulatory, or other interactions with "small" molecules and the like. Identification of tissue-enriched unique structures in RNA is another objective of the present invention.

SUMMARY OF THE INVENTION

Applicants' invention is directed to methods of identifying secondary structures in eukaryotic and prokaryotic RNA molecules termed "molecular interaction sites." Molecular interaction sites are small, preferably less than 50 nucleotides, alternatively less than 30 nucleotides, independently folded, functional subdomains contained within a larger RNA molecule. Applicants' methods preferably comprise a family of integrated processes that analyze nucleic acid, preferably RNA, sequences and predict their structure and function. Applicants' methods preferably comprise processes that execute subroutines in sequence, where the results of one process are used to trigger a specific course of action or provide numerical or other input to other steps. Preferably, there are decision points in the processes where the paths taken are determined by expert processes that make decisions without detailed, real-time human intervention. Automation of the analysis of RNA sequences provides the ability to identify regulatory sites at the rate that RNA sequences become available from genomic sequence databases and otherwise. The invention can be used, for example, to identify molecular interaction sites in connection with central nervous system (CNS) disease, metabolic disease, pain, degenerative diseases of aging, cancer, inflammatory disease, cardiovascular disease and many other conditions. Applicants' invention can also be used, for example, to identify molecular interaction sites, which are absent from eukaryotes, particularly humans, which can serves as sites for "small" molecule binding with concomitant modulation, either augmenting or diminishing, of the RNA of prokaryotic organisms. Human toxicity can, thus, be avoided in the treatment of viral, bacterial or parasitic disease.

The present invention preferably identifies molecular interaction sites in a target nucleic acid by comparing the nucleotide sequence of the target nucleic acid with the nucleotide sequences of a plurality of nucleic acids from different taxonomic species, identifying at least one sequence region which is effectively conserved among the plurality of nucleic acids and the target nucleic acid, determining whether the conserved region has secondary structure, and, for conserved regions having secondary structure, identifying the secondary structures.

The present invention is also directed to databases relating to molecular interaction sites, in eukaryotic and prokaryotic RNA. The databases are obtained by comparing the nucleotide sequence of the target nucleic acid with the nucleotide sequences

of a plurality of nucleic acids from different taxonomic species, identifying at least one sequence region which is conserved among the plurality of nucleic acids and the target nucleic acid, determining whether the conserved region has secondary structure, and for the conserved regions having secondary structure, identifying the secondary structures, and compiling a group of such secondary structures.

The present invention is also directed to oligonucleotides comprising a molecular interaction site that is present in the RNA of a selected organism and in the RNA of at least one additional organism, wherein the molecular interaction site serves as a binding site for at least one molecule which, when bound to the molecular interaction site, modulates the expression of the RNA in the selected organism.

The present invention is also directed to an oligonucleotide comprising a molecular interaction site that is present in prokaryotic RNA and in at least one additional prokaryotic RNA, wherein the molecular interaction site serves as a binding site for at least one molecule, when bound to the molecular interaction site, modulates the expression of the prokaryotic RNA.

The present invention also concerns pharmaceutical compositions comprising an oligonucleotide having a molecular interaction site that is present in prokaryotic RNA and in at least one additional prokaryotic RNA, wherein the molecular interaction site serves as a binding site for at least one "small" molecule. Such molecule, when bound to the molecular interaction site, modulates the expression of the prokaryotic RNA. A pharmaceutical carrier is also preferably included.

The present invention also provides pharmaceutical compositions comprising an oligonucleotide comprising a molecular interaction site that is present in the RNA of a selected organism and in the RNA of at least one additional organism. The molecular interaction site serves as a binding site for at least one molecule that, when bound to the molecular interaction site, modulates the expression of the RNA in the selected organism, and a pharmaceutical carrier.

Ultimately, the methods of the present invention identify the physical structures present in a target nucleic acid which are of great importance to an organism in which the nucleic acid is present. Such structures - called molecular interaction sites - are capable of interacting with molecular species to modify the nature or effect of the nucleic acid.

This may be exploited therapeutically as will be appreciated by persons skilled in the art. Such structures may also be found in the nucleic acid of organisms having great importance in agriculture, pollution control, industrial biochemistry, and otherwise. Accordingly, pesticides, herbicides, fungicides, industrial organisms such as yeast, bacteria, viruses, and the like, and biocatalytic systems may be benefitted hereby.

While there are a number of ways to characterize binding between molecular interaction sites and ligands, such as for example, organic compounds, preferred methodologies are described in, for example, U.S. Serial Numbers 09/076,440, 09/076,405, 09/076,447, 09/076,206, 09/076,214, and 09/076,404, each of which was filed on May 12, 1998 and each assigned to the assignee of this invention. All of the foregoing applications are incorporated by reference herein in their entirety.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 illustrates a flowchart comprising one preferred set of method steps for identifying molecular interaction sites in eukaryotic and prokaryotic RNA.

Figure 2 is a flowchart describing a preferred set of procedures in the Find Neighbors And Assemble ESTBlast protocol.

Figure 3 is a flowchart describing preferred steps in the BlastParse protocol.

Figure 4 is a flowchart describing preferred steps in the Q-Compare protocol.

Figures 5A, 5B, 5C and 5D illustrate flowcharts describing preferred steps in the CompareOverWins protocol.

Figure 6 is representative scatter plot of an interspecies sequence comparison between mouse and human for a ferritin RNA.

Figure 7 shows an example of self complementation analysis of a single sequence.

Figure 8 shows an overlay of self-complementarity plots of certain orthologs, and selection for the most repetitive pattern in each, resulting in a minimal number of possible folded configurations as depicted in the diagonal strings of blocks.

Figure 9 shows an exemplary descriptor.

Figure 10 shows a set of e-value scores for ferritin.

Figure 11 is a representative scatter plot of an interspecies sequence comparison between human and trout for a ferritin RNA.

Figure 12 is representative scatter plot of an interspecies sequence comparison between human and chicken for a ferritin RNA.

Figure 13 shows a representative lookup table used in Q-compare or CompareOverWins.

Figure 14 shows a representative block diagram of a program called RevComp.

Figure 15 shows a representative flow chart showing preferred steps of a preferred database search strategy for ortholog finding.

Figure 16 shows a representative Hovergen family tree for ferritin species classification.

Figure 17 shows a representative Hovergen family tree for ferritin mammalian orders classification.

Figure 18 shows a representative flow scheme showing preferred steps for a preferred SEALS strategy.

Figure 19 shows a representative plot showing regions of sequence similarity between human and mouse ferritin 5' UTR.

Figure 20 represents a genetic map showing a conserved iron response element in the 5' UTR of mouse and human ferritin.

Figure 21 shows a representative plot showing regions of sequence similarity between human and trout ferritin 5' UTR.

Figure 22 shows a representative plot showing regions of sequence similarity between human and chicken ferritin 5' UTR.

Figure 23 shows a representative Align Hits view of ferritin 5' UTR.

Figure 24 shows a representative Clustal Alignment of ferritin 5' UTR.

Figure 25 shows representative flow scheme showing preferred steps for a preferred Structure Predictor strategy.

Figure 26 shows a representative reverse complement matrix for ferritin 5' UTR.

Figure 27 shows a representative Dome structure view of ferritin 5' UTR structure.

Figure 28 shows a representative structure drawing of ferritin 5' UTR.

Figure 29 shows a representative Hovergen family tree for histone.

Figure 30 shows a representative Hovergen family tree showing vertebrate classification for histone.

Figure 31 shows a representative Align Hits view of histone 3' UTR.

Figure 32 shows a representative Clustal Alignment for histone 3' UTR.

Figure 33 shows a representative reverse complement matrix for histone 3' UTR.

Figure 34 shows a representative Dome structure view structure of histone 3' UTR.

Figure 35 shows a representative structure drawing for histone 3' UTR.

Figure 36 shows a representative structure drawing of region 2 of vimentin 3' UTR.

Figure 37 shows a representative Hovergen family tree for vimentin.

Figure 38 shows a representative Align Hits view of vimentin 3' UTR.

Figure 39 shows a representative Clustal Alignment of region 1 of vimentin 3' UTR.

Figure 40 shows a representative Dome structure view of region 1 of vimentin 3' UTR.

Figure 41 shows a representative structure drawing of region 1 of vimentin 3' UTR.

Figure 42 shows a structure proposed by Zehner *et al.* for vimentin 3' UTR.

Figure 43 shows a representative Clustal Alignment of region 2 of vimentin 3' UTR.

Figure 44 shows a representative Dome structure view of region 2 of vimentin 3' UTR.

Figure 45 shows a representative Hovergen family tree of transferrin receptor.

Figure 46 shows a representative Align Hits view of region 1 of transferrin receptor 3' UTR.

Figure 47 shows a representative Clustal Alignment of region 1 of transferrin receptor 3' UTR.

Figure 48 shows a representative Dome structure view of region 1 of transferrin receptor 3' UTR.

Figure 49 shows a representative structure drawing of region 1 of transferrin receptor 3' UTR.

Figure 50 shows a representative Align Hits view of region 2 of transferrin receptor 3' UTR.

Figure 51 shows a representative Clustal Alignment of region 2 of transferrin receptor 3' UTR.

Figure 52 shows a representative Dome structure view of region 2 of transferrin receptor 3' UTR.

Figure 53 shows a representative structure drawing of region 2 of transferrin receptor 3' UTR.

Figure 54 shows a representative Align Hits view of region 3 of transferrin receptor 3' UTR.

Figure 55 shows a representative Clustal Alignment of region 3 of transferrin receptor 3' UTR.

Figure 56 shows a representative Dome structure view of region 3 of transferrin receptor 3' UTR.

Figure 57 shows a representative structure drawing of region 3 of transferrin receptor 3' UTR.

Figure 58 shows a representative Align Hits view of region 4 of transferrin receptor 3' UTR.

Figure 59 shows a representative Clustal Alignment of region 4 of transferrin receptor 3' UTR.

Figure 60 shows a representative Dome structure view of region 4 of transferrin receptor 3' UTR.

Figure 61 shows a representative structure drawing of region 4 of transferrin receptor 3' UTR.

Figure 62 shows a representative Align Hits view of region 5 of transferrin receptor 3' UTR.

Figure 63 shows a representative Clustal Alignment of region 5 of transferrin receptor 3' UTR.

Figure 64 shows a representative Dome structure view of region 5 of transferrin receptor 3' UTR.

Figure 65 shows a representative structure drawing of region 5 of transferrin receptor 3' UTR.

Figure 66 shows a representative mass-spec structure probe analysis of region 1 of ornithine decarboxylase 3' UTR.

Figure 67 shows a representative Clustal Alignment of region 1 of ornithine decarboxylase 3' UTR.

Figure 68 shows a representative Hovergen family tree of ornithine decarboxylase 3' UTR.

Figure 69 shows a representative Hovergen family tree of vertebrate ornithine decarboxylase 3' UTR.

Figure 70 shows a representative Align Hits view of ornithine decarboxylase 3' UTR.

Figure 71 shows a representative reverse complement matrix of region 1 of ornithine decarboxylase 3' UTR.

Figure 72 shows a representative Dome structure view of region 1 of ornithine decarboxylase 3' UTR.

Figure 73 shows a representative structure drawing of region 1 of ornithine decarboxylase 3' UTR.

Figure 74 shows a representative Clustal Alignment of region 2 of ornithine decarboxylase 3' UTR.

Figure 75 shows a representative Dome structure view of region 2 of ornithine decarboxylase 3' UTR.

Figure 76 shows a representative structure drawing of region 2 of ornithine decarboxylase 3' UTR.

Figure 77 shows a representative Hovergen family tree of interleukin-2 (IL-2).

Figure 78 shows a representative Align Hits view of IL-2 3' UTR.

Figure 79 shows a representative Clustal Alignment of region 1 of IL-2 3' UTR.

Figure 80 shows a representative Dome structure view of region 1 of IL-2 3' UTR.

Figure 81 shows a representative structure drawing of region 1 of IL-2 3' UTR.

Figure 82 shows a representative Clustal Alignment of region 2 of IL-2 3' UTR.

Figure 83 shows a representative Dome structure view of region 2 of IL-2 3' UTR.

Figure 84 shows a representative structure drawing of region 2 of IL-2 3' UTR.

Figure 85 shows a representative Align Hits view of IL-2 3' UTR.

Figure 86 shows a representative Clustal Alignment of region 3 of IL-2 3' UTR.

Figure 87 shows a representative Dome structure view of region 3 of IL-2 3' UTR.

Figure 88 shows a representative structure drawing of region 3 of IL-2 3' UTR.

Figure 89 shows a representative Hovergen family tree of interleukin-4 (IL-4).

Figure 90 shows a representative Align Hits view of IL-4 5' UTR.

Figure 91 shows a representative Clustal Alignment of IL-4 5' UTR.

Figure 92 shows a representative Dome structure view of IL-4 5' UTR.

Figure 93 shows a representative structure drawing of IL-4 5' UTR.

Figure 94 shows a representative Align Hits view of IL-4 3' UTR.

Figure 95 shows a representative Clustal Alignment of IL-4 3' UTR.

Figure 96 shows a representative Dome structure view of IL-4 3' UTR.

Figure 97 shows a representative structure drawing of IL-4 3' UTR.

The present invention is directed to methods of identifying particular structural elements in eukaryotic and prokaryotic nucleic acid, especially RNA molecules, which will interact with other molecules to effect modulation of the RNA. "Modulation" refers to augmenting or diminishing RNA activity or expression. A preferred embodiment of the present invention is outlined in flowchart form in Figure 1. The structural elements in

eukaryotes and prokaryotes are referred to as "molecular interaction sites." These elements contain secondary structure, that is, have three-dimensional form capable of undergoing interaction with "small" molecules and otherwise, and are expected to serve as sites for interacting with "small" molecules, oligomers such as oligonucleotides, and other compounds in therapeutic and other applications.

Referring to Figure 1, preferred steps for identifying molecular interaction sites in target nucleic acids are shown in the flow diagram. The nucleotide sequence of the target nucleic acid is compared with the nucleotide sequences of a plurality of nucleic acids from different taxonomic species, 10. The target nucleic acid may be present in eukaryotic cells or prokaryotic cells, the target nucleic acid may be bacterial or viral as well as belonging to a "higher" organism such as human. Any type of nucleic acid can serve as a target nucleic acid. Preferred target nucleic acids include, but are not limited to, messenger RNA (mRNA), pre-messenger RNA (pre-mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), or small nuclear RNA (snRNA). Initial selection of a particular target nucleic acid can be based upon any functional criteria. Nucleic acids known to be important during inflammation, cardiovascular disease, pain, cancer, arthritis, trauma, obesity, Huntingtons, neurological disorders, or other diseases or disorders, for example, are exemplary target nucleic acids.

Nucleic acids known to be involved in pathogenic genomes such as, for example, bacterial, viral and yeast genomes are exemplary prokaryotic nucleic acid targets. Pathogenic bacteria, viruses and yeast are well known to those skilled in the art. Exemplary nucleic acid targets are shown in Table 1. Applicants' invention, however, is not limited to the targets shown in Table 1 and it is to be understood that the present invention is believed to be quite general.

Table 1: Exemplary RNA Targets

Protein	RNA Target	GenBank #	Therapeutic
46 kD protein	3'-UTR stemloop in vimentin mRNA	X56134	cancer
unknown-cGMP regulated	5'-UTR of Asialoglycoprotein receptor mRNA	m10058	cancer

unknown	unknown	m11025	unknown
unknown insulin regulated protein	3'-UTR of E-selectin mRNA	unknown	inflammation
30 kD protein	3'-UTR of lipoprotein lipase mRNA	m15856	obesity
unknown	5'-UTR of NR2A subunit of NMDA receptor	U09002	trauma, paid, AD
histone binding protein (HBP)	3'-UTR of histone mRNA + paralogs	x57129	cancer
unknown	3'-UTR of p53 mRNA	x02469	cancer
p53	5'-UTR of mdm2 oncogene mRNA	u39736	cancer
unknown	5'-UTR of interleukin 1 type receptor (IL-1R1)	m27492	inflammation
none	5'-UTR of muscle acylphosphatase mRNA	x84195	musculoskeletal disease
ribosomal proteins	5'-UTR of c-myc in multiple myeloma	V00568	cancer
unknown	5'-UTR of Huntingtons disease gene		Huntingtons
unknown	5'-UTR of angiotensin AT	p30556	cardiovascular disease
unknown	zip code sequence in ARC mRNA	d87468	unknown
L-4	5'-UTR of L4 ribosomal protein	d23660	cancer
L-32	5'-UTR of L32 ribosomal protein	x03342	cancer
unknown	TCTP, translationally controlled tumor protein	x16064	cancer
unknown	3'-UTR of B-F1-ATPase	d00022	cancer
PU family of proteins, FBF binding factor	3'-UTR of fem-3 in *C. elegans*	X64962	unknown

unknown	3'-UTR of myocyte enhancer factor 2 MEF2A	x68505	metabolic
unknown	5'-UTR of glucose transporter mRNA GLUT1	k03195	diabetes
48 kD reticulocyte protein	3'-UTR of 15-lipoxygenase	M23892	inflammation
La proetin	5'-UTR of ribosomal RNA proteins		cancer
unknown	translational regulation of IL-2	S82692	inflammation
unknown	3'-UTR of CaMKIIa mRNA in neurons	u81554	CNS
bicoid (bcd)	BRE 3'-UTR fragment mRNA encoding cad protein	M21069	under development
48/50 kD protein	3'-UTR structure protamines 1	Y00443	cancer
translin (human) TB-RBP (mouse)	protamine 1 mRNA (human testes specific)	Y00443	cancer
translin (human) TB-RBP (mouse)	protamine 2 mRNA	X07862	unknown
translin (human) TB-RBP (mouse)	transition protein mRNA	x14474	cancer
translin (human) TB-RBP (mouse)	Tau mRNA	m13577	cancer
translin (human) TB-RBP (mouse)	myelin basic protein mRNA	x07948	cancer
p75	3'-UTR of ribonucleotide reductase R2	x59618	cancer
39 kD poly C protein	alpha globin	v00493	cancer
unknown	beta protein	v00497	metabolic
human teratocarcinoma protein p40	Line-1 mRNA		cancer, metabolic

RPL32	5'-UTR hairpin structure in RPL32		cancer
Y-box proteins	family of transcription factor mRNAs with a Y-box sequence		cancer
telomerase protein	telomerase RNA	AF015950	cancer
ferritin, transferrin	IREs, internal loops in mRNA encoding ferritin and transferrin		inflammation
ribosomal proteins	5'-UTR of PDGF2/c-sis mRNA	M12873	inflammation
zip code for localization	3'-UTR of beta actin		cancer
unknown insulin regulated protein	5'-UTR of ornithine decarboxylase mRNA	x55362	cancer
ribosomal proteins	ornithine decarboxylase antizyme		cancer
unknown	FGF-5		inflammation
DFR protein factor	3'-UTR TGE elements in the human oncogene GLI	X07384	cancer
DFR protein factor	3'-UTR tra-2 of *C. elegans*		unknown
viral capsid protein	3'-UTR of alfalfa mosaic virus RNA3		unknown
unknown	BRE Bruno response element in 3'-UTR of drosophila oskar mRNA		cancer
unknown	NRE nanose response element		cancer
unknown	repeated element		inflammation
U1A RDB protein	U1 snRNA		inflammation
CD40		X60592	inflammation
IGF-R		X04434 M24599	inflammation

A1 adenosine receptor		X68485	cardiovascular
B7-1		M27533	inflammation
B7-2			inflammation
cyclophilin B		M60857 M60457 M63573	inflammation
cyclophilin C		S71018	transplantation
FKBP51			transplantation
Th1 cytokines IFN γ			inflammation
Th1 cytokines IL-12		U03187	inflammation
NF-kappa B			cancer
ICAM-1		X06990	inflammation
L-selectin		X16150	inflammation
VCAM-1		M30257	inflammation
Alpha 4 integrin		X16983 X15356	inflammation
Beta 7		U34971	inflammation
MadCAM-1		U43628	inflammation
PECAM-1		M28526	inflammation
LFA-1		Y00796	inflammation
TACE			inflammation
LFA-3		X06296 Y00636	inflammation
CD-18			inflammation
ICAM-3		X69819	inflammation
ICAM-2		X15606	inflammation
CD11a		M87662	inflammation
protein kinase C-α			cancer

protein kinase C-β		X52479	cancer
protein kinase C-δ			cancer
protein kinase C-ε		Z22521	cancer
protein kinase C-h		X65293	cancer
protein kinase C-m		M55284	cancer
protein kinase C-ζ			cancer
unknown		Z15108	unknown
unknown	ornithine decarboxylase mRNA	X55362	cancer
unknown	IL-2 mRNA	X01586	inflammation
unknown	IL-4	M13982	inflammation

Additional nucleic acid targets may be determined independently or can be selected from publicly available prokaryotic and eukaryotic genetic databases known to those skilled in the art. Preferred databases include, for example, Online Mendelian Inheritance in Man (OMIM), the Cancer Genome Anatomy Project (CGAP), GenBank, EMBL, PIR, SWISS-PROT, and the like. OMIM, which is a database of genetic mutations associated with disease, was developed, in part, for the National Center for Biotechnology Information (NCBI). OMIM can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/Omim/. CGAP, which is an interdisciplinary program to establish the information and technological tools required to decipher the molecular anatomy of a cancer cell. CGAP can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/ncicgap/. Some of these databases may contain complete or partial nucleotide sequences. In addition, nucleic acid targets can also be selected from private genetic databases. Alternatively, nucleic acid targets can be selected from available publications or can be determined especially for use in connection with the present invention.

After a nucleic acid target is selected or provided, the nucleotide sequence of the nucleic acid target is determined and then compared to the nucleotide sequences of a plurality of nucleic acids from different taxonomic species. In one embodiment of the invention, the nucleotide sequence of the nucleic acid target is determined by scanning at least

one genetic database or is identified in available publications. Preferred databases known and available to those skilled in the art include, for example, the Expressed Gene Anatomy Database (EGAD) and Unigene-Homo Sapiens database (Unigene), GenBank, and the like. EGAD contains a non-redundant set of human transcript (HT) sequences and can be accessed through the Internet at, for example, http://www.tigr.org/tdb/egad/egad.html. Unigene is a system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each Unigene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

In addition, Unigene contains hundreds of thousands of novel expressed sequence tag (EST) sequences. Unigene can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/UniGene/. These databases can be used in connection with searching programs such as, for example, Entrez, which is known and available to those skilled in the art, and the like. Entrez can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/Entrez/. Preferably, the most complete nucleic acid sequence representation available from various databases is used. The GenBank database, which is known and available to those skilled in the art, can also be used to obtain the most complete nucleotide sequence. GenBank is the NIH genetic sequence database and is an annotated collection of all publicly available DNA sequences. GenBank is described in, for example, *Nuc. Acids Res.*, **1998**, *26*, 1-7, which is incorporated herein by reference in its entirety, and can be accessed by those skilled in the art through the Internet at, for example, http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html. Alternatively, partial nucleotide sequences of nucleic acid targets can be used when a complete nucleotide sequence is not available.

In another embodiment of the present invention, the nucleotide sequence of the nucleic acid target is determined by assembling a plurality of overlapping expressed sequence tags (ESTs). The EST database (dbEST), which is known and available to those skilled in the art, comprises approximately one million different human mRNA sequences comprising from about 500 to 1000 nucleotides, and various numbers of ESTs from a number of different organisms. dbEST can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/dbEST/index.html. These sequences are derived from a cloning

strategy that uses cDNA expression clones for genome sequencing. ESTs have applications in the discovery of new genes, mapping of genomes, and identification of coding regions in genomic sequences. Another important feature of EST sequence information that is becoming rapidly available is tissue-specific gene expression data. This can be extremely useful in targeting selective gene(s) for therapeutic intervention. Since EST sequences are relatively short, they must be assembled in order to provide a complete sequence. Because every available clone is sequenced, it results in a number of overlapping regions being reported in the database.

Assembly of overlapping ESTs extended along both the 5' and 3' directions results in a full-length "virtual transcript." The resultant virtual transcript may represent an already characterized nucleic acid or may be a novel nucleic acid with no known biological function. The Institute for Genomic Research (TIGR) Human Genome Index (HGI) database, which is known and available to those skilled in the art, contains a list of human transcripts. TIGR can be accessed through the Internet at, for example, http://www.tigr.org/. The transcripts were generated in this manner using TIGR-Assembler, an engine to build virtual transcripts and which is known and available to those skilled in the art. TIGR-Assembler is a tool for assembling large sets of overlapping sequence data such as ESTs, BACs, or small genomes, and can be used to assemble eukaryotic or prokaryotic sequences. TIGR-Assembler is described in, for example, Sutton, *et al.*, *Genome Science & Tech.*, **1995**, *1*, 9-19, which is incorporated herein by reference in its entirety, and can be accessed through the Internet at, for example, ftp://ftp.tigr.org/pub/software/TIGR assembler. In addition, GLAXO-MRC, which is known and available to those skilled in the art, is another protocol for constructing virtual transcripts. In addition, "Find Neighbors and Assemble EST Blast" protocol, which runs on a UNIX platform, has been developed by Applicants to construct virtual transcripts. Preferred steps in the Find Neighbors and Assemble EST Blast protocol is described in the flowchart set forth in Figure 2. PHRAP is used for sequence assembly within Find Neighbors and Assemble EST Blast. PHRAP can be accessed through the Internet at, for example, http://chimera.biotech.washington.edu/uwgc/tools/phrap.htm. One skilled in the art can construct source code to carry out the preferred steps set forth in Figure 2.

The nucleotide sequence of the nucleic acid target is compared to the nucleotide sequences of a plurality of nucleic acids from different taxonomic species. A

plurality of nucleic acids from different taxonomic species, and the nucleotide sequences thereof, can be found in genetic databases, from available publications, or can be determined especially for use in connection with the present invention. In one embodiment of the invention, the nucleic acid target is compared to the nucleotide sequences of a plurality of nucleic acids from different taxonomic species by performing a sequence similarity search, an ortholog search, or both, such searches being known to persons of ordinary skill in the art.

The result of a sequence similarity search is a plurality of nucleic acids having at least a portion of their nucleotide sequences which are homologous to at least an 8 to 20 nucleotide region of the target nucleic acid, referred to as the window region. Preferably, the plurality of nucleotide sequences comprise at least one portion which is at least 60% homologous to any window region of the target nucleic acid. More preferably, the homology is at least 70%. More preferably, the homology is at least 80%. Most preferably, the homology is at least 90%. For example, the window size, the portion of the target nucleotide to which the plurality of sequences are compared, can be from about 8 to about 20, preferably 10 - 15, most preferably about 11 - 12, contiguous nucleotides. The window size can be adjusted accordingly. A plurality of nucleic acids from different taxonomic species is then preferably compared to each likely window in the target nucleic acid until all portions of the plurality of sequences is compared to the windows of the target nucleic acid. Sequences of the plurality of nucleic acids from different taxonomic species which have portions which are at least 60%, preferably at least 70%, more preferably at least 80%, or most preferably at least 90% homologous to any window sequence of the target nucleic acid are considered as likely homologous sequences.

Sequence similarity searches can be performed manually or by using several available computer programs known to those skilled in the art. Preferably, Blast and Smith-Waterman algorithms, which are available and known to those skilled in the art, and the like can be used. Blast is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein sequence databases. Blast can be accessed through the Internet at, for example, http://www.ncbi.nlm.nih.gov/BLAST/. The GCG Package provides a local version of Blast that can be used either with public domain databases or with any locally available searchable database. GCG Package v.9.0 is a commercially available software package that contains over 100 interrelated software programs that enables analysis

of sequences by editing, mapping, comparing and aligning them. Other programs included in the GCG Package include, for example, programs which facilitate RNA secondary structure predictions, nucleic acid fragment assembly, and evolutionary analysis. In addition, the most prominent genetic databases (GenBank, EMBL, PIR, and SWISS-PROT) are distributed along with the GCG Package and are fully accessible with the database searching and manipulation programs. GCG can be accessed through the Internet at, for example, http://www.gcg.com/. Fetch is a tool available in GCG that can get annotated GenBank records based on accession numbers and is similar to Entrez. Another sequence similarity search can be performed with GeneWorld and GeneThesaurus from Pangea. GeneWorld 2.5 is an automated, flexible, high-throughput application for analysis of polynucleotide and protein sequences. GeneWorld allows for automatic analysis and annotations of sequences. Like GCG, GeneWorld incorporates several tools for homology searching, gene finding, multiple sequence alignment, secondary structure prediction, and motif identification. GeneThesaurus 1.0tm is a sequence and annotation data subscription service providing information from multiple sources, providing a relational data model for public and local data.

Another alternative sequence similarity search can be performed, for example, by BlastParse. BlastParse is a PERL script running on a UNIX platform that automates the strategy described above. BlastParse takes a list of target accession numbers of interest and takes each one through the preferred processes described in the flowchart set forth in Figure 3. BlastParse parses all the GenBank fields into "tab-delimited" text that can then be saved in a "relational database" format for easier search and analysis, which provides flexibility. The end result is a series of completely parsed GenBank records that can be easily sorted, filtered, and queried against, as well as an annotations-relational database.

Another toolkit capable of doing sequence similarity searching and data manipulation is SEALS, also from NCBI. This tool set is written in perl and C and can run on any computer platform that supports these languages. It is available for download, for example, at: http://www.ncbi.nlm.nih.gov/Walker/SEALS/. This toolkit provides access to Blast2 or gapped blast. It also includes a tool called tax_collector which, in conjuction with a tool called tax_break, parses the output of Blast2 and returns the identifier of the sequence most homologous to the query sequence for each species present. Another useful tool is feature2fasta which extracts sequence fragments from an input sequence based on the

annotation. An exemplary use for this tool is to create sequence files containing the 5' untranslated region of a cDNA sequence.

Preferably, the plurality of nucleic acids from different taxonomic species which have homology to the target nucleic acid, as described above in the sequence similarity search, are further delineated so as to find orthologs of the target nucleic acid therein. An ortholog is a term defined in gene classification to refer to two genes in widely divergent organisms that have sequence similarity, and perform similar functions within the context of the organism. In contrast, paralogs are genes within a species that occur due to gene duplication, but have evolved new functions, and are also referred to as isotypes. Optionally, paralog searches can also be performed. By performing an ortholog search, an exhaustive list of homologous sequences from diverse organisms is obtained. Subsequently, these sequences are analyzed to select the best representative sequence that fits the criteria for being an ortholog. An ortholog search can be performed by programs available to those skilled in the art including, for example, Compare. Preferably, an ortholog search is performed with access to complete and parsed GenBank annotations for each of the sequences. Currently, the records obtained from GenBank are "flat-files", and are not ideally suited for automated analysis. Preferably, the ortholog search is performed using a Q-Compare program. Preferred steps of the Q-Compare protocol are described in the flowchart set forth in Figure 4. The Blast Results-Relation database, depicted in Figure 3, and the Annotations-Relational database, depicted in Figure 3, are used in the Q-Compare protocol, which results in a list of ortholog sequences to compare in the interspecies sequence comparisons programs described below.

The above-described similarity searches provide results based on cut-off values, referred to as e-scores. E-scores represent the probability of a random sequence match within a given window of nucleotides. The lower the e-score, the better the match. One skilled in the art is familiar with e-scores. The user defines the e-value cut-off depending upon the stringency, or degree of homology desired, as described above. In embodiments of the invention where prokaryotic molecular interaction sites are identified, it is preferred that any homologous nucleotide sequences that are identified be non-human.

In another embodiment of the invention, the sequences required are obtained by searching ortholog databases. One such database is Hovergen, which is a curated database

of vertebrate orthologs. Ortholog sets may be exported from this database and used as is, or used as seeds for further sequence similarity searches as described above. Further searches may be desired, for example, to find invertebrate orthologs. Hovergen can be downloaded, for example, at: ftp://pbil.univ-lyon1.fr/pub/hovergen/. A database of prokaryotic orthologs, COGS, is available and can be used interactively on the internet, for example at: http://www.ncbi.nlm.nih.gov/COG/.

In another embodiment of the present invention, the nucleotide sequences of a plurality of nucleic acids from different taxonomic species are compared to the nucleotide sequence of the target nucleic acid by performing a sequence similarity search using dbEST, or the like, and constructing virtual transcripts. Using EST information is useful for two distinct reasons. First, the ability to identify orthologs for human genes in evolutionarily distinct organisms in GenBank database is limited. As more effort is directed towards identifying ESTs from these evolutionarily distinct organisms, dbEST is likely to be a better source of ortholog information.

Second, the attempt to sequence human genome is less than 10 % complete. Thus, it is likely that the human dbEST will provide more information for identifying primary targets as the sequence of the human genome nears completion. EST sequences are short and need to be assembled to be used. Preferably, a sequence similarity search is performed using Smith-Waterman algorithms, as described above, under high stringency against dbEST excluding human sequences. Because dbEST contains sequencing errors, including insertions and deletions, in order to accurately search for new sequences, the search method used should allow for these gaps. Because every available clone is sequenced, it results in a number of overlapping regions being reported in the database. A full-length or partial "virtual transcript" for non-human RNAs is constructed by a process whereby overlapping EST sequences are extended along both the 5' and 3' directions, until a "full-length" transcript is obtained. In another embodiment of the invention, a chimeric virtual transcript is constructed.

The resultant virtual transcript may represent an already characterized RNA molecule or could be a novel RNA molecule with no known biological function. As described above, TIGR HGI database makes available an engine to build virtual transcripts called TIGR-Assembler. GLAXO-MRC and GeneWorld from Pangea provide for construction of

virtual transcripts as well. As described above, Find Neighbors and Assemble EST Blast can also be used to build virtual transcripts.

Referring to Figure 1, after the orthologs or virtual transcripts described above are obtained through either the sequence similarity search or the ortholog search, at least one sequence region which is conserved among the plurality of nucleic acids from different taxonomic species and the target nucleic acid is identified, 20. Interspecies sequence comparisons can be performed using numerous computer programs which are available and known to those skilled in the art. Preferably, interspecies sequence comparison is performed using Compare, which is available and known to those skilled in the art. Compare is a GCG tool that allows pair-wise comparisons of sequences using a window/stringency criterion. Compare produces an output file containing points where matches of specified quality are found. These can be plotted with another GCG tool, DotPlot.

Alternatively, the identification of a conserved sequence region is performed by interspecies sequence comparisons using the ortholog sequences generated from Q-Compare in combination with CompareOverWins, as described above. Preferably, the list of sequences to compare, *i.e.*, the ortholog sequences, generated from Q-Compare, as described in Figure 4, is entered into the CompareOverWins algorithm. Preferred steps in the CompareOverWins are described in Figures 5A, 5B, and 5C. Preferably, interspecies sequence comparisons are performed by a pair-wise sequence comparison in which a query sequence is slid over a window on the master target sequence. Preferably, the window is from about 9 to about 99 contiguous nucleotides.

Sequence homology between the window sequence of the target nucleic acid and the query sequence of any of the plurality of nucleic acid sequences obtained as described above, is preferably at least 60%, more preferably at least 70%, more preferably at least 80%, and most preferably at least 90%. The most preferable method of choosing the threshold is to have the computer automatically try all thresholds from 50% to 100% and choose a threshold based a metric provided by the user. One such metric is to pick the threshold such that exactly n hits are returned, where n is usually set to 3. This process is repeated until every base on the query nucleic acid, which is a member of the plurality of nucleic acids described above, has been compared to every base on the master target sequence. The resulting scoring matrix can be plotted as a scatter plot. Based on the match density at a given location, there

may be no dots, isolated dots, or a set of dots so close together that they appear as a line. The presence of lines, however small, indicates primary sequence homology. A representative scatter plot of such interspecies sequence comparison is depicted in Figure 6. Sequence conservation within nucleic acid molecules, particularly the UTRs of RNA, in divergent species is likely to be an indicator of conserved regulatory elements that are also likely to have a secondary structure. The results of the interspecies sequence comparison can be analyzed using MS Excel and visual basic tools in an entirely automated manner as known to those skilled in the art.

Referring to Figure 1, after at least one region that is conserved between the nucleotide sequence of the nucleic acid target and the plurality of nucleic acids from different taxonomic species, preferably via the orthologs, is identified, the conserved region is analyzed to determine whether it contains secondary structure, 30. Determining whether the identified conserved regions contain secondary structure can be performed by a number of procedures known to those skilled in the art. Determination of secondary structure is preferably performed by self complementarity comparison, alignment and covariance analysis, secondary structure prediction, or a combination thereof.

In one embodiment of the invention, secondary structure analysis is performed by alignment and covariance analysis. Numerous protocols for alignment and covariance analysis are known to those skilled in the art. Preferably, alignment is performed by ClustalW, which is available and known to those skilled in the art. ClustalW is a tool for multiple sequence alignment that, although not a part of GCG, can be added as an extension of the existing GCG tool set and used with local sequences. ClustalW can be accessed through the Internet at, for example, http://dot.imgen.bcm.tmc.edu:9331/multi-align/Options/clustalw.html. ClustalW is also described in Thompson, *et al.*, *Nuc. Acids Res.*, **1994**, *22*, 4673-4680, which is incorporated herein by reference in its entirety. These processes can be scripted to automatically use conserved UTR regions identified in earlier steps. Seqed, a UNIX command line interface available and known to those skilled in the art, allows extraction of selected local regions from a larger sequence. Multiple sequences from many different species can be clustered and aligned for further analysis.

In a preferred embodiment of the invention, the output of all possible pair-wise CompareOverWindows comparisons are compiled and aligned to a reference sequence using

a program called AlignHits. A diagram of the operation of this program is given in Figure 5D. This program could be reproduced by one skilled in the art. A preferred purpose of this program is to map all hits made in pair-wise comparisons back to the position on a reference sequence. This method combining CompareOverWindows and AlignHits provides more local alignments (over 20-100 bases) than any other algorithm. This local alignment is required for the structure finding routines described later such as covariation or RevComp. This algorithm writes a fasta file of aligned sequences. As shown, the algorithm does not correct single base insertions or deletions. This is usually accomplished by putting the output through ClustalW described elsewhere. It is important to differentiate this from using ClustalW by itself, without CompareOverWindows and AlignHits.

Covariation is a process of using phylogenetic analysis of primary sequence information for consensus secondary structure prediction. Covariation is described in the following references, each of which is incorporated herein by reference in their entirety: Gutell, *et al.*, "Comparative Sequence Analysis Of Experiments Performed During Evolution" In Ribosomal RNA Group I Introns, Green, Ed., Austin:Landes, **1996**; Gautheret, *et al.*, *Nuc. Acids Res.*, **1997**, *25*, 1559-1564; Gautheret, *et al.*, *RNA*, **1995**, *1*, 807-814; Lodmell, *et al.*, *Proc. Natl. Acad. Sci. USA*, **1995**, *92*, 10555-10559; Gautheret, *et al.*, *J. Mol. Biol.*, **1995**, *248*, 27-43; Gutell, *Nuc. Acids Res.*, **1994**, *22*, 3502-3517; Gutell, *Nuc. Acids Res.*, **1993**, *21*, 3055-3074; Gutell, *Nuc. Acids Res.*, **1993**, *21*, 3051-3054; Woese, *Proc. Natl. Acad. Sci. USA*, **1989**, *86*, 3119-3122; and Woese, *et al.*, *Nuc. Acids Res.*, **1980**, *8*, 2275-2293. Preferably, covariance software is used for covariance analysis. Preferably, Covariation, a set of programs for the comparative analysis of RNA structure from sequence alignments, is used. Covariation uses phylogenetic analysis of primary sequence information for consensus secondary structure prediction. Covariation can be obtained through the Internet at, for example, http://www.mbio.ncsu.edu/RNaseP/info/programs/programs.html. A complete description of a version of the program has been published (Brown, J. W. 1991 Phylogenetic analysis of RNA structure on the Macintosh computer. CABIOS7:391-393). The current version is v4.1, which can perform various types of covariation analysis from RNA sequence alignments, including standard covariation analysis, the identification of compensatory base-changes, and mutual information analysis. The program is well-documented and comes with extensive example files. It is compiled as a stand-alone program; it does not require

Hypercard (although a much smaller 'stack' version is included). This program will run in any Macintosh environment running MacOS v7.1 or higher. Faster processor machines (68040 or PowerPC) is suggested for mutual information analysis or the analysis of large sequence alignments.

In another embodiment of the invention, secondary structure analysis is performed by secondary structure prediction. There are a number of algorithms that predict RNA secondary structures based on thermodynamic parameters and energy calculations. Preferably, secondary structure prediction is performed using either M-fold or RNA Structure 2.52. M-fold can be accessed through the Internet at, for example, http://www.ibc.wustl.edu/-zuker/ma/form2.cgi or can be downloaded for local use on UNIX platforms. M-fold is also available as a part of GCG package. RNA Structure 2.52 is a windows adaptation of the M-fold algorithm and can be accessed through the Internet at, for example, http://128.151.176.70/RNAstructure.html.

In another embodiment of the invention, secondary structure analysis is performed by self complementarity comparison. Preferably, self complementarity comparison is performed using Compare, described above. More preferably, Compare can be modified to expand the pairing matrix to account for G-U or U-G basepairs in addition to the conventional Watson-Crick G-C/C-G or A-U/U-A pairs. Such a modified Compare program (modified Compare) begins by predicting all possible base-pairings within a given sequence. As described above, a small but conserved region, preferably a UTR, is identified based on primary sequence comparison of a series of orthologs. In modified Compare, each of these sequences is compared to its own reverse complement. Figure 7 depicts an exemplary self complementarity analysis. Allowable base-pairings include Watson-Crick A-U, G-C pairing and non-canonical G-U pairing. An overlay of such self complementarity plots of all available orthologs, and selection for the most repetitive pattern in each, results in a minimal number of possible folded configurations. Figure 8 shows an exemplary overlay. These overlays can then used in conjunction with additional constraints, including those imposed by energy considerations described above, to deduce the most likely secondary structure.

In another preferred embodiment of the invention, the output of AlignHits is read by a program called RevComp. A block diagram of this program is shown in Figure 14. This program could be reproduced by one skilled in the art. A preferred purpose of this

program is to use base pairing rules and ortholog evolution to predict RNA secondary structure. RNA secondary structures are composed of single stranded regions and base paired regions, called stems. Since structure conserved by evolution is searched, the most probable stem for a given alignment of ortholog sequences is the one which could be formed by the most sequences. Possible stem formation or base pairing rules is determined by, for example, analyzing base pairing statistics of stems which have been determined by other techniques such as NMR. The output of RevComp is a sorted list of possible structures, ranked by the percentage of ortholog set member sequences which could form this structure. Because this approach uses a percentage threshold approach, it is insensitive to noise sequences. Noise sequences are those that either not true orthologs, or sequences that made it into the output of AlignHits due to high sequence homology even though they do not represent an example of the structure which is searched. A very similar algorithm is implemented using Visual basic for Applications (VBA) and Microsoft Excel to be run on PCs, to generate the reverse complement matrix view for the given set of sequences.

A result of the secondary structure analysis described above, whether performed by alignment and covariance, self complementarity analysis, secondary structure predictions, such as using M-fold or otherwise, is the identification of secondary structure in the conserved regions among the target nucleic acid and the plurality of nucleic acids from different taxonomic species, 40. Exemplary secondary structures that may be identified include, but are not limited to, bulges, loops, stems, hairpins, knots, triple interacts, cloverleafs, or helices, or a combination thereof. Alternatively, new secondary structures may be identified.

In another embodiment of the invention, once the secondary structure of the conserved region has been identified, as described above, at least one structural motif for the conserved region having secondary structure is identified. These structural motifs correspond to the identified secondary structures described above. For example, analysis of secondary structure by self complementation may provide one type of secondary structure, whereas analysis by M-fold may provide another secondary structure. All the possible secondary structures identified by secondary structure analysis described above are, thus, represented by a family of structural motifs.

Once the secondary structure(s) of the target nucleic acids, as well as the secondary structures of nucleic acids from different taxonomic species, have been identified, further nucleic acids can be identified by searching on the basis of structure, rather than by primary nucleotide sequence, as described above. Additional nucleic acids which have secondary structure similar or identical to the secondary structure found as described above can be identified by constructing a family of descriptor elements for the structural motifs described above, and identifying other nucleic acids having secondary structures corresponding to the descriptor elements. The combination of any or all of the nucleic acids having secondary structure can be compiled into a database. The entire process can be repeated with a different target nucleic acid to generate a plurality of different secondary structure groups which can be compiled into the database. Thus, databases of molecular interaction sites can be compiled by performing by the invention described herein.

After the hypothetical structure motifs are determined from the secondary structure analysis described above, a family of structure descriptor elements is constructed. Preferably, the structural motifs described above are converted into a family of descriptor elements. An exemplary descriptor element is shown in Figure 9. One skilled in the art is familiar with construction of descriptors. Structure descriptors are described in, for example, Laferriere, *et al.*, *Comput. Appl. Biosci.*, **1994**, *10*, 211-212, incorporated herein by reference in its entirety. A different structure descriptor element is constructed for each of the structural motifs identified from the secondary structure analysis. Briefly, the secondary structure is converted to a generic text string, such as shown in Figure 9. For novel motifs, further biochemical analysis such as chemical mapping or mutagenesis may be needed to confirm structure predictions. Descriptor elements may be defined to have various stringency.

For example, referring to Figure 9, the region termed H1, which comprises the first region of the stem, can be described as NNN:NNN, which contemplates any complementary base pairing including G-C, C-G, A-U, and U-A. The H1 region may also be designated so as to include only C-G or A-U, etc., base pairing. In addition, the descriptor elements can be defined to allow for a wobble. Thus, descriptor elements can be defined to have any level of stringency desired by the user. Applicants' invention, thus, is also directed to a database comprising different descriptor elements.

After a family of structure descriptor elements is constructed, nucleic acids having secondary structure which correspond to the structure descriptor elements are identified. Preferably, nucleic acids having secondary structure which correspond to the structure descriptor elements are identified by searching at least one database, performing clustering and analysis, identifying orthologs, or a combination thereof. Thus, the identified nucleic acids have secondary structure which falls within the scope of the secondary structure defined by the descriptor elements. Thus, the identified nucleic acids have secondary structure identical to nearly identical, depending on the stringency of the descriptor elements, to the target nucleic acid.

In one embodiment of the invention, nucleic acids having secondary structure which correspond to the structure descriptor elements are identified by searching at least one database. Any genetic database can be searched. Preferably, the database is a UTR database, which is a compilation of the untranslated regions in messenger RNAs. A UTR database is accessible through the Internet at, for example, ftp://area.ba.cnr.it/pub/embnet/database/utr/. Preferably the database is searched using a computer program, such as, for example, Rnamot, a UNIX-based motif searching tool available from Daniel Gautheret. Each "new" sequence that has the same motif is then queried against public domain databases to identify additional sequences. Results are analyzed for recurrence of pattern in UTRs of these additional ortholog sequences, as described below, and a database of RNA secondary structures is built. One skilled in the art is familiar with Rnamot. Briefly, Rnamot takes a descriptor string, such as the one shown in Figure 9, and searches any Fasta format database for possible matches. Descriptors can be very specific, to match exact nucleotide(s), or can have built-in degeneracy. Lengths of the stem and loop can also be specified. Single stranded loop regions can have a variable length. G-U pairings are allowed and can be specified as a wobble parameter. Allowable mismatches can also be included in the descriptor definition. Functional significance is assigned to the motifs if their biological role is known based on previous analysis. Known regulatory regions such as Iron Response Element have been found using this technique (see, Example 1 below). In embodiments of the invention in which a database containing prokaryotic molecular interaction sites is compiled, it is preferable to refrain from searching human sequences or, alternatively, discarding human sequences when found.

In another embodiment of the invention, the nucleic acids identified by searching databases such as, for example, searching a UTR database using Rnamot, are clustered and analyzed so as to determine their location within the genome. The results provided by Rnamot simply identify sequences containing the secondary structure but do not give any indication as to the location of the sequence in the genome. Clustering and analysis is preferably performed with ClustalW, as described above.

In another embodiment of the invention, after clustering and analysis is performed as described above, orthologs are identified as described above. However, in contrast to the orthologs identified above, which were solely identified on the basis of their primary nucleotide sequences, these new orthologous sequences are identified on the basis of structure using the nucleic acids identified using Rnamot. Identification of orthologs is preferably performed by BlastParse or Q-Compare, as described above. In embodiments of the invention in which a database containing prokaryotic molecular interaction sites is compiled, it is preferable to refrain from finding human orthologs or, alternatively, discarding human orthologs when found.

After nucleic acids having secondary structures which correspond to the structure descriptor elements are identified, any or all of the nucleotide sequences can be compiled into a database by standard compiling protocols known to those skilled in the art. One database may contain eukaryotic molecule interaction sites and another database may contain prokaryotic molecule interaction sites.

The present invention is also directed to oligonucleotides comprising a molecular interaction site that is present in the RNA of a selected organism and in the RNA of at least one preferably several additional organisms. The nucleotide sequence of the oligonucleotide is selected to provide the secondary structure of the molecular interaction sites described above. The nucleotide sequence of the oligonucleotide is preferably the nucleotide sequence of the target nucleic acids described above. Alternatively, the nucleotide sequence is preferably the nucleotide sequence of nucleic acid from a plurality of different taxonomic species which also contain the molecular interaction site. The molecular interaction site serves as a binding site for at least one molecule which, when bound to the molecular interaction site, modulates the expression of the RNA in the selected organism.

The present invention is also directed to oligonucleotides comprising a molecular interaction site that is present in a prokaryotic RNA and in at least one additional prokaryotic RNA, wherein the molecular interaction site serves as a binding site for at least one molecule which, when bound to the molecular interaction site, modulates the expression of the prokaryotic RNA. The additional organism is selected from all eukaryotic and prokaryotic organisms and cells but is not the same organism as the selected organism. Oligonucleotides, and modifications thereof, are well known to those skilled in the art. The oligonucleotides of the invention can be used, for example, as research reagents to detect, for example, naturally occurring molecules which bind the molecular interaction sites. The oligonucleotides of the invention can also be used as decoys to compete with naturally-occurring molecular interaction sites within a cell for research, diagnostic and therapeutic applications. Molecules which bind to the molecular interaction site modulate, either by augmenting or diminishing, the expression of the RNA. The oligonucleotides can also be used in agricultural, industrial and other applications.

The present invention is also directed to pharmaceutical compositions comprising the oligonucleotides described above in combination with a pharmaceutical carrier. A "pharmaceutical carrier" is a pharmaceutically acceptable solvent, diluent, suspending agent or any other pharmacologically inert vehicle for delivering one or more nucleic acids to an animal, and are well known to those skilled in the art. The carrier may be liquid or solid and is selected, with the planned manner of administration in mind, so as to provide for the desired bulk, consistency, *etc.*, when combined with the other components of a pharmaceutical composition. Typical pharmaceutical carriers include, but are not limited to, binding agents (*e.g.*, pregelatinised maize starch, polyvinylpyrrolidone or hydroxypropyl methylcellulose, *etc.*); fillers (*e.g.*, lactose and other sugars, microcrystalline cellulose, pectin, gelatin, calcium sulfate, ethyl cellulose, polyacrylates or calcium hydrogen phosphate, *etc.*); lubricants (*e.g.*, magnesium stearate, talc, silica, colloidal silicon dioxide, stearic acid, metallic stearates, hydrogenated vegetable oils, corn starch, polyethylene glycols, sodium benzoate, sodium acetate, *etc.*); disintegrates (*e.g.*, starch, sodium starch glycolate, *etc.*); or wetting agents (*e.g.*, sodium lauryl sulphate, *etc.*).

The following examples are meant to be exemplary of the preferred embodiments of the invention and are not meant to be limiting.

EXAMPLES

Example 1: The Iron Responsive Element

1. Selecting RNA Target

To illustrate the strategy for identifying small molecule interaction sites, the iron responsive element (IRE) in the mRNA encoded by the human ferritin gene is identified. The IRE is a typical example of an RNA structural element that is used to control the level of translation of mRNAs associated with iron metabolism. The structure of the IRE was recently determined using NMR spectroscopy. In addition, NMR analysis of IRE structure is described in Gdaniec, *et al.*, *Biochem.*, **1998**, *37*, 1505-1512 and Addess, *et al.*, *J. Mol. Biol.*, **1997**, *274*, 72-83. The IRE is an RNA element of approximately 30 nucleotides that folds into a hairpin structure and binds a specific protein. Because this structure has been so well studied and it known to appear in the mRNA of many species, it serves an excellent example of how Applicants' methodology works.

2. Determining Nucleotide Sequence of the RNA Target

The human mRNA sequence for ferritin is used as the initial mRNA of interest or master sequence. The ferritin protein sequence is also used in the analysis, particularly in the initial steps used to find related sequences. In the case of human ferritin gene, the best input is the full length annotated mRNA and protein sequence obtained from UNIGENE. However, for many genes of interest the same level of detailed information is not available. In these cases, alternative sources of master sequence information is obtained from sources such as, for example, GenBank, TIGR, dbEST division of GenBank or from sequence information obtained from private laboratories. Applicants' methods work using any level of input sequence information, but requires fewer steps with a high quality annotated input sequence.

3. Identifying Similar Sequences

An early step in the process is to use the master sequence (nucleotide or protein) to find and rank related sequences in the database (orthologs and paralogs). Sequence similarity search algorithms are used for this purpose. All sequence similarity algorithms calculate a quantitative measure of similarity for each result compared with the master

sequence. An example of a quantitative result is an E-value obtained from the Blast algorithm. The E-values for a blast search of the non-redundant GenBank database using ferritin mRNA as the query sequence illustrates the use of quantitative analysis of sequence similarity searches. The E-value is the probability that a match between a query sequence and a database sequence occurs due to random chance. Therefore, the lower an E-value the more likely that two sequences are truly related. A plot of the lowest E-value scores for ferritin is shown in Figure 10. Sequences that meet the cutoff criteria are selected for more detailed comparisons according to a set of rules described below. Since an objective of the sequence similarity search to find distantly related orthologs and paralogs, it is preferable that the cutoff criteria not be too stringent, or the target of the search may be excluded.

4. Identification of Conserved Regions

Identification of conserved regions is performed by pairwise sequence comparisons using Q-Compare in conjunction with CompareOverWins. Conservation of structure between genes with related function from different species is a major indication that can be used to find good drug binding sites. Conserved structure can be identified by using distantly related sequences and piecing together the remnants of conserved sequence combining it with an analysis of potential structure. Sequence comparisons are made between pairs of mRNAs from different species using Q-compare that can identify traces of sequence conservation from even very divergent organisms. Q-compare, in conjunction with CompareOverWins, compares every region of each sequence by sliding one sequence over the other from end to end and measuring the number of matches in a window of a specific size.

When the human mRNA and mouse mRNA sequences for ferritin, which each contain an IRE in the 5'-UTR, are analyzed in this manner, a plot showing the regions of sequence similarity is produced, as shown in Figure 19. Pairwise analysis of the human and mouse ferritin mRNA sequences illustrate several important aspects of this type of analysis. Regions of each mRNA that encode the amino acid sequence have the highest degree of similarity, while the untranslated regions are less similar. In Figure 19, the location of the IRE is indicated. In both the human and mouse ferritin mRNAs the IREs are located in the extreme 5' end of each mRNA. This demonstrates an important point -- the sequence

conservation in the region of the IRE structure does not stand out against the background of sequence similarity between the human and mouse ferritin sequences. In contrast, in the comparison of human and trout (Figure 11) or human and chicken (Figure 12) ferritin mRNAs, the IREs can be immediately identified. This is because the sequence of the UTRs between human and trout or human and chicken are separated by greater evolutionarily distance than human and mouse, which is logical in view of the evolutionary distance that separates humans from birds and fish compared with other mammals. Comparing the human sequence to that of birds and fish is informative because the natural drift due to evolution has allowed many sequence changes in the UTRs. However, the IRE sequences are more constrained because they form an important structure. Thus, they stand out better and can be more readily identified.

The same principle applies when comparing the trout and chicken ferritin sequences to each other. While both are separated from humans by hundreds of millions of years of evolution, they are also well separated from each other. This illustrates another important tactic used in the present invention -- comparison of two non-human RNA sequences can be used to find a regulatory RNA structure without having the actual human sequence. The non-human comparison work can actually direct one skilled in the art where to look to find a human counterpart as a potential drug target.

Evolutionary distances can be used to decide which sequences not to compare as well as which to compare. As with the human and mouse, comparison of trout and salmon are less informative because the species are too close and the IRE does not stand out above the UTR background. Comparison of human and Drosophia ferritin mRNA sequences fail to find the IREs in either species, even though they are present. This is because the sequence of the IREs between humans and Drosophila have diverged even though the structure is conserved. However, if the Drosophila and mosquito ferritin mRNAs are compared, the IREs are identified, again illustrating that the human sequence need not be in hand to identify a regulatory element relevant to drug discovery in humans.

The software used in the present invention makes the decision whether or not to compare sequences pairwise using a lookup table based upon the evolutionary distances between species. An example of a small lookup table using the examples described above is shown in Figure 13. The lookup table in the present invention includes all species that have

sequences deposited in GenBank. Q-Compare in conjunction with CompareOverWins decides which sequences to compare pairwise.

5. Identification Of Secondary Structure

Sets of sequences that show evidence of conservation in orthologs and paralogs or other related genes are analyzed for the ability to form internal structure. This is accomplished by analyzing each sequence in a matrix where the seqeunce is plotted 5' to 3' on the X axis and its reverse complement is plotted 5' to 3' on the Y axis, such as in, for example, self-complementary analysis. Matches that correspond to potential intramolecular base pairs are scored according to a table of values. When the human ferritin IRE sequence is analyzed in this fashion, the diagonals indicate potential self- complementary regions. Each of the 13 IRE sequences described in this example were analyzed in the same fashion. While each of the sequences can form a variety of different structures, the structure most likely to occur is one common to all the sequences. By superimposing the plots of all 13 individual sequences (see, Figure 8), the potential structure common to all the sequences is deduced.

Example 2: The Iron Responsive Element (Method B)

2. Determining Nucleotide Sequence of the RNA Target

The human mRNA sequence for ferritin is used as the initial mRNA of interest or master sequence. The ferritin protein sequence is also used in the analysis, particularly in the initial steps used to find related sequences. In the case of human ferritin gene, the best input is the full length annotated mRNA (gi507251) and protein sequence obtained from UNIGENE. However, for many genes of interest the same level of detailed information is not available. In these cases, alternative sources of master sequence information is obtained from sources such as, for example, Hovergen and GenBank. The present methods work using any level of input sequence information, but requires fewer steps with a high quality annotated input sequence.

3. Identifying Similar Sequences

An alternate, and preferred, approach to finding orthologs is the use of Hovergen database and query tools that have been described in Duret, *et al.*, *Nuc. Acids Res.*, **1994**, *22*, 2360-2365, which is incorporated herein by reference in its entirety.

The use of Hovergen to identify related sequences is shown in Figure 16 (tree classification at the species level) and Figure 17 (classification at the order level). Sequences corresponding to each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding region was extracted using SEALS and COWX, as shown in Figure 18.

4. Identification of Conserved Regions

The IRE sequences are more constrained because they form an important structure. Thus, they stand out better and can be more readily identified even in closely related sequences. However, for this to work for any gene, the compare algorithm has been rewritten (*see*, Figures 5A-C). This new tool, CompareOverWins, allows a dynamic selection of both the range of window sizes, as well the hit threshold. This algorithm needs as its input parsed and separated 5' and 3' UTR sequences. We use tools available within the Seals genome analysis package described earlier to achieve this. Figure 18 describes the steps involved.

To identify the iron responseve element using the methods described herein, the compare over widows algorithm was used and the results visualized using AlignHits (Figure 5D for the algorithm). Representative results are shown in Figure 23. In addition to optimizing the thresholding, CompareOverWins also extracts the sequence corresponding to the hits. ClustalW (version 1.74) was used on the extracted sequences to create a locally gapped alignment (*see*, Figure 24). A representative flow scheme for this approach is shown in Figure 25.

5. Identification Of Secondary Structure

Sets of sequences that show evidence of conservation in orthologs and paralogs or other related genes are analyzed for the ability to form internal structure. This is accomplished by analyzing each sequence in a matrix where the seqeunce is plotted 5' to 3'

on the X axis and its complement is plotted 5' to 3' on the Y axis, such as in, for example, self-complementary analysis. Matches that correspond to potential intramolecular base pairs are scored according to a table of values. When the human ferritin IRE sequence is analyzed in this fashion, the diagonals indicate potential self- complementary regions. Each of the 13 IRE sequences described in this example were analyzed in the same fashion. While each of the sequences can form a variety of different structures, the structure most likely to occur is one common to all the sequences. By superimposing the plots of all 13 individual sequences (*see,* Figure 26), the potential structure common to all the sequences is deduced.

The above scheme has been implemented algorithmically into a program called RevComp (*see*, Figure 14). RevComp creates a sorted list of all the structures. Representative results can be viewed either as a "dome" ouptut (*see,* Figure 27) or as a "connect" or "ct" file which can be used in one of many RNA structure viewing programs (RNAStructure, RNAViz, etc.). A representative example of such a structure drawing is shown in Figure 28.

Example 3: Histone

Histone 3'UTR represents another classic stem-loop structure that has been studied extensively (*EMBO*, **1997**, *16*, 769). At the post-transcriptional level, the stem-loop structure in the 3' untranslated region of the histone mRNA has been shown to be very important. Son, *Saenghwahak Nyusu*, **1993**, *13*, 64-70. The analysis shown below describes the use of this known structure to validate the strategy and methods described herein.

Figures 29 and 30 represent phylogenetic tree outputs for all Histone orthologs in Hovergen database. Each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding regions were extracted and compared using SEALS and COWX as described earlier (*see,* Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions (*see*, Figure 31). One such region is shown encircled. The sequences corresponding to the region of interest was extracted from all species for alignment with CLUSTAL W (1.74). Following extraction of sequence information from Align Hits, CLUSTAL W (1.74) was used to provide multiple sequence

alignment shown (*see*, Figure 32). Each of the putative hit sequence was analyzed for the ability to form internal structure. This was accomplished by analyzing each sequence in a matrix where the sequence was plotted 5' to 3' on the X axis and its complement is plotted 5' to 3' on the Y axis. Base-pairs along the diagonals indicate potential self-complementary regions that can form secondary structures. Figure 33 shows a representative reverse complement matrix. Figure 34 shows a representative sequence alignment in a dome format showing potential stem formation between the base pairs. Following conversion of the dome format file to a ct file, RNA Structure 3.21 is used to visualize the structure (*see*, Figure 35).

Example 4: Vimentin

Vimentin is an intermediate filament protein whose 3'UTR is highly conserved between species. Previous studies by Zehner *et al.*, (*Nuc. Acids Res.*, **1997**, *25*, 3362-3370) has shown that a proposed a complex stem-loop structure contained within this region may be important for vimentin mRNA functions such as mRNA localization. The same region was identified using the present analysis, thus validating the present approach. In addition, based on the analyses described herein, a second stem-loop structure that occurs downstream of the previously proposed structure that may have a role in regulating vimentin fuction as well has been identified (*see*, Figure 36).

A representative phylogenetic tree output for all Vimentin orthologs in Hovergen database is shown in Figure 37. Each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding regions were extracted and compared using SEALS and COWX as described earlier (*see*, Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions. Two such regions appear, and were used for subsequent analyses (*see*, Figure 38). Following extraction of sequence information from Align Hits for region 1, CLUSTAL W was used to provide multiple sequence alignment shown (*see*, Figure 39). Potential stem formation between base pairs is given above the sequence alignment in a dome format is shown in Figure 40. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure (*see*, Figure 41). This structure is very similar to the one proposed by Zehner *et al.* (*see*, Figure

42). Zehner *et al.* presented a detailed chemical analysis of their proposed structure for the minimal binding domain in the 3' UTR of Vimentin. This analysis included cleavage with single-strand-specific (ChS or T1) or double-strand-specific (V1) nucleases as well as after exposure to lead acetate.

Following extraction of sequence information from Align Hits for region 2, CLUSTAL W was used to provide multiple sequence alignment shown in Figure 43. The potential stem formation between base pairs in region 2 is given above the sequence alignment in a dome format (*see*, Figure 44). Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure for the region 2 (*see*, Figure 36).

Example 5: Transferrin Receptor

Similar to regulation of ferritin (Examples 1 and 2), another known function of the IRE is in the regulation of transferrin receptor. Five IREs have been identified in the 3' UTRs of known transferring receptor mRNAs. Kuhn *et al.*, *EMBO J.*, **1987**, *6*, 1287-93 and Casey *et al.*, *Science*, **1988**, *240*, 924-928, each of which is incorporated herein by reference in its entirety. All 5 IREs have been shown to interact with iron regulatory proteins (IRP) independently. The present techniques were applied to identify these conserved elements in transferrin receptors.

A representative phylogenetic tree output for all Transferrin receptor orthologs in Hovergen database is shown in Figure 45. Each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding region were extracted and compared using SEALS and COWX as described earlier (*see*, Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions as shown in Figure 46. This can be seen where a vertical line intersects a series of horizontal lines representing sequence information from a set of species. This region between base pairs 920 to 990 in the 3 prime UTR of transferrin receptor was extracted from all species for alignment with CLUSTAL W (1.74).

Following extraction of sequence information from Align Hits for region 1, CLUSTAL W (1.74) was used to provide multiple sequence alignment as shown in Figure 47. A representative potential stem formation between base pairs is given above the sequence

alignment in a dome format as shown in Figure 48. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure (*see*, Figure 49). This can be seen where a vertical line intersects a series of horizontal lines representing sequence information from a set of species. This region between base pairs 990 to 1050 in the 3 prime UTR of transferrin receptor was extracted from all species for alignment with CLUSTAL W (1.74) (*see*, Figure 50).

Following extraction of sequence information from Align Hits for region 2, CLUSTAL W (1.74) was used to provide multiple sequence alignment as shown in Figure 51. Potential stem formation between base pairs is given above the sequence alignment in a dome format as shown in Figure 52. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure as shown in Figure 53. Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions. This can be seen where a vertical line intersects a series of horizontal lines representing sequence information from a set of species. This region between base pairs 1372 to 1423 in the 3 prime UTR of transferrin receptor was extracted from all species for alignment with CLUSTAL W (1.74) (*see*, Figure 54).

Following extraction of sequence information from Align Hits for region 3, CLUSTAL W (1.Ex.34) was used to provide multiple sequence alignment as shown in Figure 55. Potential stem formation between base pairs is given above the sequence alignment in a dome format as shown in Figure 56. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure as shown in Figure 57. Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions. This can be seen where a vertical line intersects a series of horizontal lines representing sequence information from a set of species. This region between base pairs 1439 to 1479 in the 3 prime UTR of transferrin receptor was extracted from all species for alignment with CLUSTAL W (1.74) (*see*, Figure 58).

Following extraction of sequence information from Align Hits for region 4, CLUSTAL W (1.Ex.34) was used to provide multiple sequence alignment as shown in Figure 59. Potential stem formation between base pairs is given above the sequence alignment in a dome format is shown in Figure 60. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure as shown in Figure 61. Following

extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions. This can be seen where a vertical line intersects a series of horizontal lines representing sequence information from a set of species. This region between base pairs 1479 to 1542 in the 3 prime UTR of transferrin receptor was extracted from all species for alignment with CLUSTAL W (1.74) (*see*, Figure 62).

Following extraction of sequence information from Align Hits for region 5, CLUSTAL W (1.Ex.34) was used to provide multiple sequence alignment as shown in Figure 63. Potential stem formation between base pairs is given above the sequence alignment in a dome format is shown in Figure 64. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure as shown in Figure 65.

Example 6: Ornithine Decarboxylase

Orinithine decarboxylase (ODC) is the first enzyme in the polyamine biosynthetic pathway. Studies have shown existence of translational regulatory elements both in the 5' and 3' untranslated regions (Grens *et al.*, *J. Biol. Chem.*, **1990**, *265*, 11810). Secondary structures have been proposed to exist in both these regions, though there is no conclusive evidence for it. The methods described herein identified two structures in the 3' UTR, as shown below. The presence of one of these structures (*see*, Figure 66) was verified using mass spectrometry probing (Griffey, *et al.*, *Proc. SPIE-Int. Soc. Opt. Eng.*, 2985 (Ultrasensitive Biochemical Diagnostics II): 82-86, which is incorporated herein by reference in its entirety). Two representative sequences that showed slight variation in their lengths (*see*, Figure 67) were made into RNA and subjected to MS structure probing. Results shown in Figure 66 confirm the presence of a stem-loop structure. Accordingly, identification of a novel secondary structure can be identified from the methods described herein, and such existence has been independently verified by structure probing.

Phylogenetic tree outputs for all Ornithine Decarboxylase orthologs in Hovergen database is shown in Figure 68 and Figure 69. Each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding region were extracted and compared using SEALS and COWX as described earlier (*see*, Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions as shown in Figure 70. Two such regions appear, and were used for subsequent analyses. Following extraction of sequence information from region1, CLUSTAL W (1.74) was used to provide multiple sequence alignment shown. Each of the putative hit sequences was analyzed for the ability to form internal structure as shown in the reverse complement matrix depicted in Figure 71. This was accomplished by analyzing each sequence in a matrix where the sequence is plotted 5' to 3' on the X axis and its complement is plotted 5' to 3' on the Y axis. Base-pairs along the diagonals indicate potential self- complementary regions that can form secondary structures. Domes view of the potential stem formation between base pairs in region 1 is given above the sequence alignment was determined using RevComp (*see*, Figure 72). RNA Structure 3.2 was used to visualize the structure (*see*, Figure 73).

Mass spectrometry analyses techniques were used to probe for structure. Figure 67 showed presence of gaps/inserts in the multiple alignment. Two representative RNAs (gi404561 and gi35135) from the alignments shown in Figure 67 were used for this experiment. Analysis of the pattern of induced fragmentation showed a very strong likelihood for base-paring along the top half of the stem-loop structure (shown inverted in the figure). This corresponds to bases 11-14 and 20-23 in 404561 or bases 8-11 and 18-21 in 35135. Bulged bases (G9 in 404561 or U22 in 35135) also showed characteristic fragmentation pattern. The bottom-half of the structure appeared to be less stable, and showed some fragmentation where our analyses had predicted base-paring. This was particularly true in the sequence 35135. This region, however, has several contiguous A-U or G-U base-pairs which tend to be less stable, and therefore have a higher probability of fragmentation.

Following extraction of sequence information from Align Hits for region 2, CLUSTAL W was used to provide multiple sequence alignment shown as shown in Figure 74. Potential stem formation between base pairs in region 2 is given above the sequence alignment in a dome format as shown in Figure 75. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure for the region 2 as shown in Figure 76.

Example 7: Interleukin-2 (IL-2)

A representative phylogenetic tree output for all IL-2 orthologs in Hovergen database is shown in Figure 77. Each of these orthologs was saved in GenBank format and grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding region were extracted and compared using SEALS and COWX as described earlier (*see*, Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions in the 3'UTR region. Two such regions appear, and were used for subsequent analyses (*see*, Figure 78). Following extraction of sequence information from Align Hits for region 1, CLUSTAL W (1.74) was used to provide multiple sequence alignment shown in Figure 79. Domes view of the potential stem formation between base pairs in region 1 is given above the sequence alignment was determined using RevComp (*see*, Figure 80). RNA Structure 3.2 was used to visualize the structure as depicted in Figure 81. Following extraction of sequence information from Align Hits for region 2, CLUSTAL W (1.74) was used to provide multiple sequence alignment shown in Figure 82. Potential stem formation between base pairs in region 2 is given above the sequence alignment in a dome format as shown in Figure 83. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure for the region 2 as shown in Figure 84.

In addition to the two regions described above, a third region, downstream of, and partially overlapping region 2, was identified using an alternate reference sequence (3087784.fa) and is shown in Figure 85. Following extraction of sequence information from Align Hits for this region, CLUSTAL W (1.74) was used to provide multiple sequence alignment shown in Figure 86. Potential stem formation between base pairs in region 3 is shown in Figure 87 above the sequence alignment in a dome format. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure for region 3 (*see*, Figure 88).

Example 8: Interleukin-4 (IL-4)

Representative phylogenetic tree output for all IL-4 orthologs in Hovergen database is shown in Figure 89. Each of these orthologs was saved in GenBank format and

grouped together in a single data file. Untranslated regions in both the 5' and 3' flanks of the coding region were extracted and compared using SEALS and COWX as described earlier (*see*, Figures 18 and 25).

Following extraction and comparison by SEALS and COWX, Align Hits was used to determine potentially interesting regions in the 5'UTR region as shown in Figure 90. Following extraction of sequence information from Align Hits for the above region, CLUSTAL W (1.74) was used to provide multiple sequence alignment shown in Figure 91. Domes view of the potential stem formation between base pairs in region is given above the sequence alignment was determined using RevComp (*see*, Figure 92). RNA Structure 3.2 was used to visualize the structure as shown in Figure 93.

Figure 94 depicts a representative Align Hits view of hits in the 3'UTR region of IL-4. Following extraction of sequence information from Align Hits for the 3' UTR region, CLUSTAL W (1.74) was used to provide multiple sequence alignment as shown in Figure 95. Potential stem formation between base pairs in region 2 is given above the sequence alignment in a dome format is shown in Figure 96. Following conversion of the dome format file to a ct file, RNA Structure 3.21 was used to visualize the structure for the region 2 (*see*, Figure 97).

What is claimed is:

1. A method of identifying molecular interaction sites in a target nucleic acid comprising:

comparing the nucleotide sequence of said target nucleic acid with the nucleotide sequences of a plurality of nucleic acids from different taxonomic species;

identifying at least one sequence region which is conserved among said plurality of nucleic acids and said target nucleic acid;

determining whether said conserved region has secondary structure; and

for said conserved region having secondary structure, identifying said secondary structure.

2. The method of claim 1 further comprising identifying at least one structural motif for said conserved region having secondary structure.

3. The method of claim 2 further comprising constructing a set of descriptor elements for said structural motif.

4. The method of claim 3 further comprising identifying further nucleic acids having secondary structures corresponding to said descriptor elements.

5. The method of claim 1 wherein said target nucleic acid is present in a eukaryotic cell.

6. The method of claim 5 wherein said target nucleic acid is selected from the group consisting of mRNA, pre-mRNA, tRNA, rRNA, and snRNA.

7. The method of claim 1 wherein said target nucleic acid is present in a prokaryotic cell.

8. The method of claim 7 wherein said target nucleic acid is RNA.

9. The method of claim 7 wherein said target nucleic acid is bacterial.

10. The method of claim 7 wherein said target nucleic acid is viral.

11. The method of claim 7 wherein said target nucleic acid is from a parasite.

12. The method of claim 1 wherein at least some nucleic acid sequence information is derived from a genetic database.

13. The method of claim 1 wherein said nucleotide sequence of said target nucleic acid is determined by assembling a plurality of expressed sequence tags.

14. The method of claim 1 further comprising comparing said target nucleic acid to paralogous nucleic acids.

15. The method of claim 1 wherein said plurality of nucleic acids from different taxonomic species is obtained by performing a sequence similarity search, an ortholog search, or a combination thereof.

16. The method of claim 1 wherein said plurality of nucleic acids from different taxonomic species is obtained by performing a sequence similarity search and constructing virtual transcripts.

17. The method of claim 1 wherein determining whether said conserved region has secondary structure is performed by self complementarity comparison, alignment and covariance analysis, secondary structure prediction, or a combination thereof.

18. The method of claim 17, wherein said secondary structure comprises at least one bulge, loop, stem, hairpin, knot, triple interact, cloverleaf, or helix.

19. The method of claim 2 wherein said structural motif is identified by performing self complementarity comparison, alignment and covariance analysis, secondary structure prediction, or a combination thereof.

20. The method of claim 3 wherein said set of descriptor elements is constructed using a descriptor database.

21. The method of claim 4 wherein said other nucleic acids having secondary structures corresponding to said descriptor elements are identified by searching at least one database, performing clustering and analysis, searching for orthologs, or a combination thereof.

22. A database containing molecular interaction sites identified by the method of claim 1.

23. The database of claim 22 containing eukaryotic molecular interaction sites.

24. The database of claim 23 containing human molecular interaction sites.

25. The database of claim 22 containing prokaryotic molecular interaction sites.

26. An oligonucleotide comprising a molecular interaction site that is present in the RNA of a selected organism and in the RNA of at least one additional organism, wherein said molecular interaction site serves as a binding site for at least one molecule that when bound to said molecular interaction site modulates the expression of said RNA in said selected organism.

27. An oligonucleotide comprising a molecular interaction site that is present in prokaryotic RNA and in at least one additional prokaryotic RNA, wherein said molecular interaction site serves as a binding site for at least one molecule that when bound to said molecular interaction site modulates the expression of said prokaryotic RNA.

28. The oligonucleotide of claim 27 wherein said molecular interaction site is not present in eukaryotic RNA.

29. The oligonucleotide of claim 27 wherein said molecular interaction site is not present in human RNA.

30. A pharmaceutical composition comprising:

an oligonucleotide comprising a molecular interaction site that is present in a prokaryotic RNA and in at least one additional prokaryotic RNA, wherein said molecular interaction site serves as a binding site for at least one molecule that when bound to said molecular interaction site modulates the expression of said prokaryotic RNA; and

a pharmaceutical carrier or diluent.

31. The pharmaceutical composition of claim 30 wherein said molecular interaction site is not present in eukaryotic RNA.

32. The pharmaceutical composition of claim 30 wherein said molecular interaction site is not present in human RNA.

33. A pharmaceutical composition comprising:

an oligonucleotide comprising a molecular interaction site that is present in the RNA of a selected organism and in the RNA of at least one additional organism, wherein said molecular interaction site serves as a binding site for at least one molecule that when bound to said molecular interaction site modulates the expression of said RNA in said selected organism; and

a pharmaceutical carrier or diluent.

34. A pharmaceutical composition comprising:

an oligonucleotide comprising a molecular interaction site present in a prokaryotic RNA, which site is not present in mamalian RNA; and

a pharmaceutical carrier or diluent.

ABSTRACT

Methods of identifying molecular interaction sites in eukaryotic and prokaryotic nucleic acids, especially RNA, are described. Secondary structural elements are identified from highly conserved sequences. Methods of preparing databases relating to such molecular interaction sites are also provided herein as are databases themselves. Therapeutic, agricultural, industrial, and other applicability results from interaction of such molecular interaction sites with "small" and other molecules.

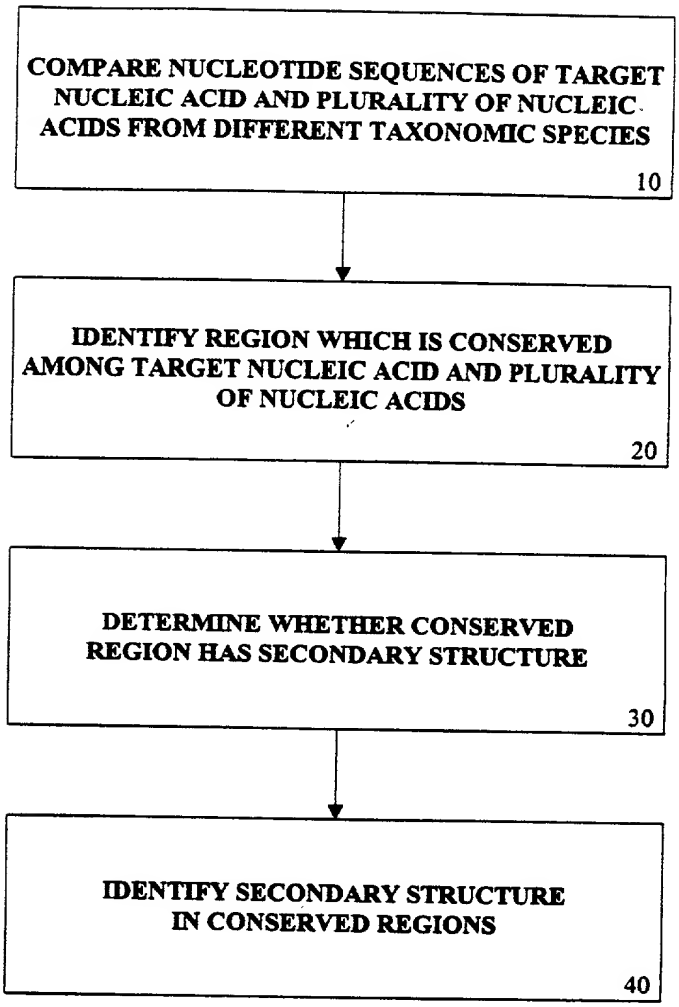

COMPARE NUCLEOTIDE SEQUENCES OF TARGET NUCLEIC ACID AND PLURALITY OF NUCLEIC ACIDS FROM DIFFERENT TAXONOMIC SPECIES
10
IDENTIFY REGION WHICH IS CONSERVED AMONG TARGET NUCLEIC ACID AND PLURALITY OF NUCLEIC ACIDS
20
DETERMINE WHETHER CONSERVED REGION HAS SECONDARY STRUCTURE
30
IDENTIFY SECONDARY STRUCTURE IN CONSERVED REGIONS
40

FIGURE 1

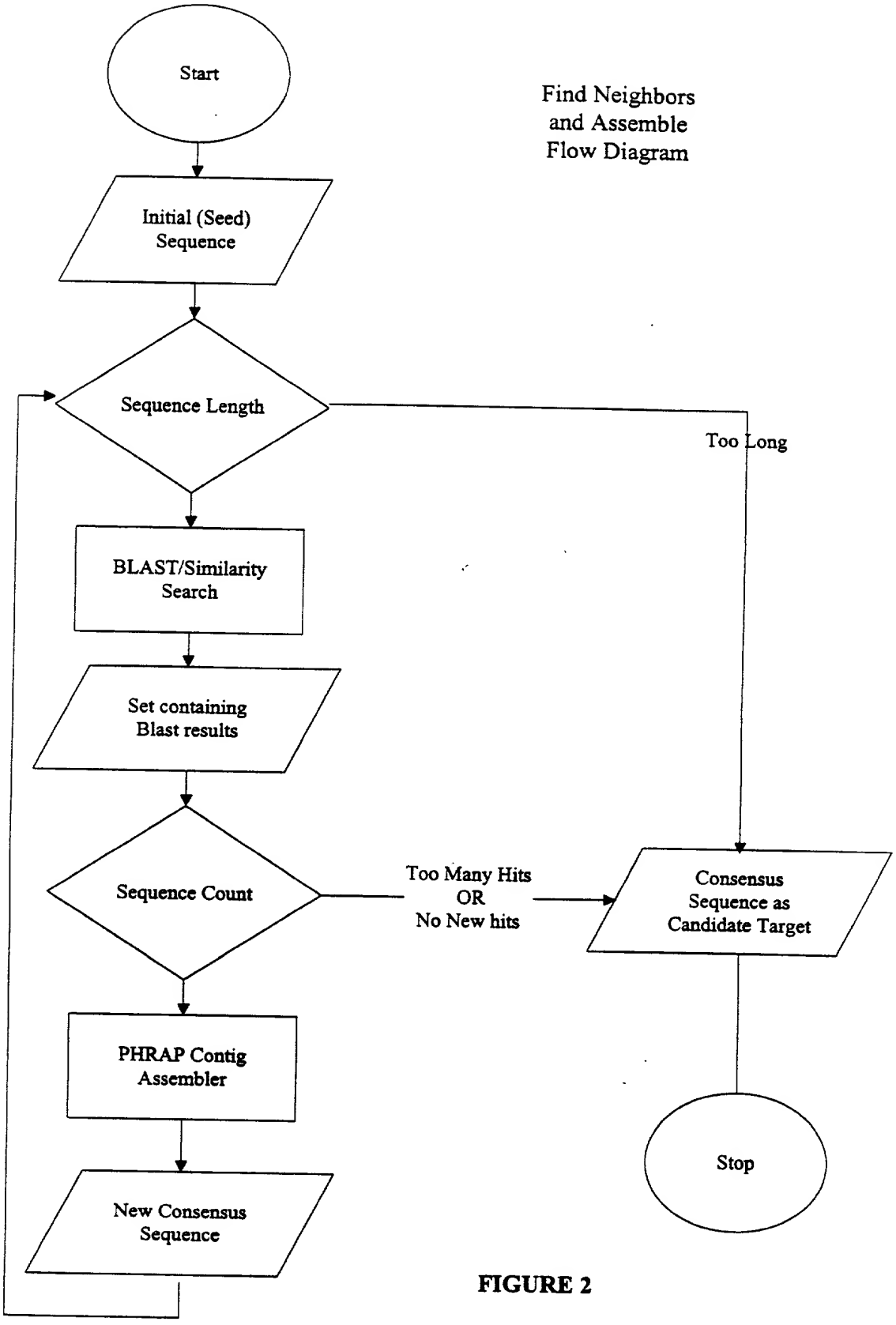

Find Neighbors
and Assemble
Flow Diagram
Start
Initial (Seed)
Sequence
Sequence Length
Too Long
BLAST/Similarity
Search
Set containing
Blast results
Sequence Count
Too Many Hits
OR
No New hits
Consensus
Sequence as
Candidate Target
PHRAP Contig
Assembler
Stop
New Consensus
Sequence

FIGURE 2

BlastParse

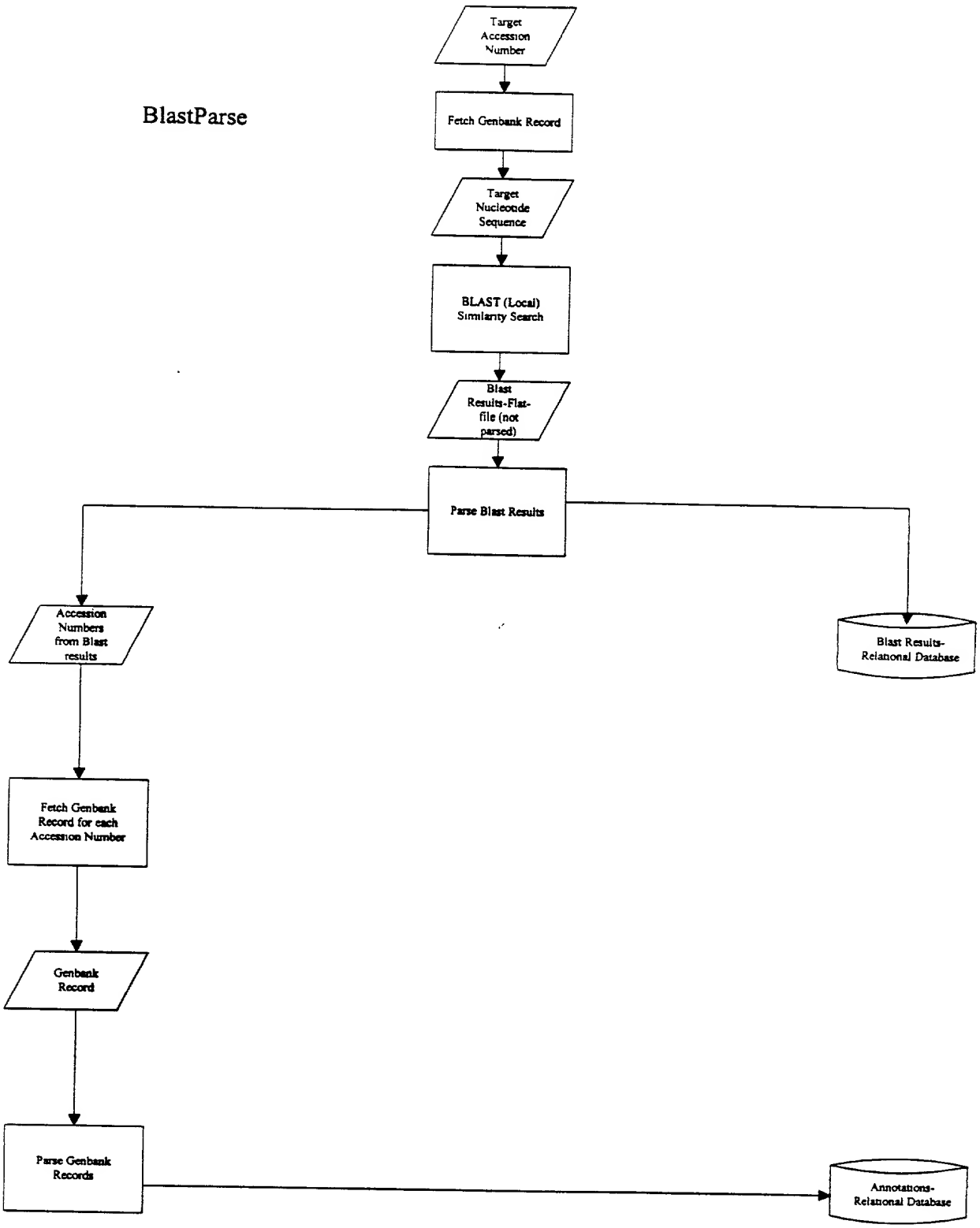


FIGURE 3

Q-Compare

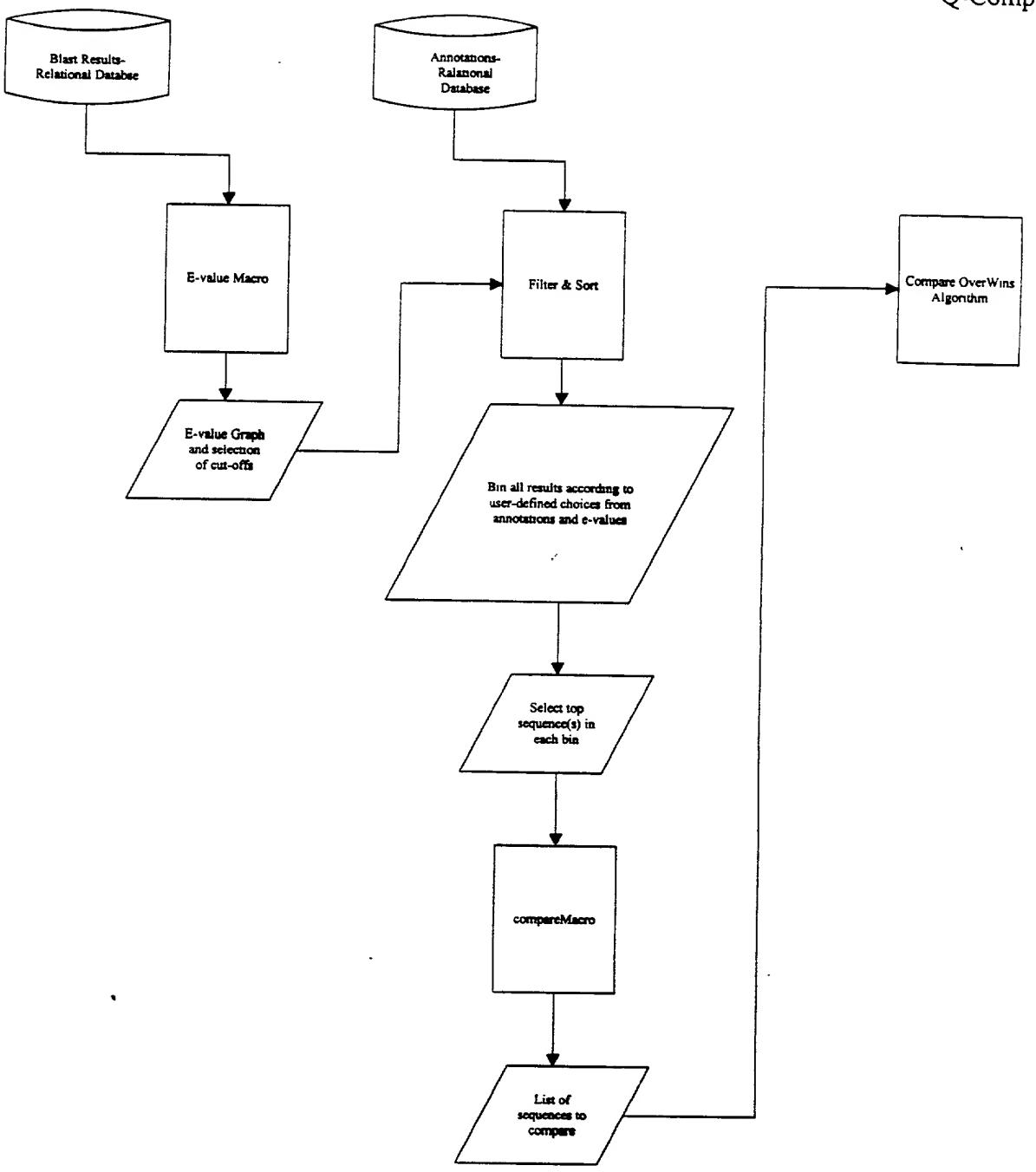


FIGURE 4

CompareOverWins Algorithm Flow Chart

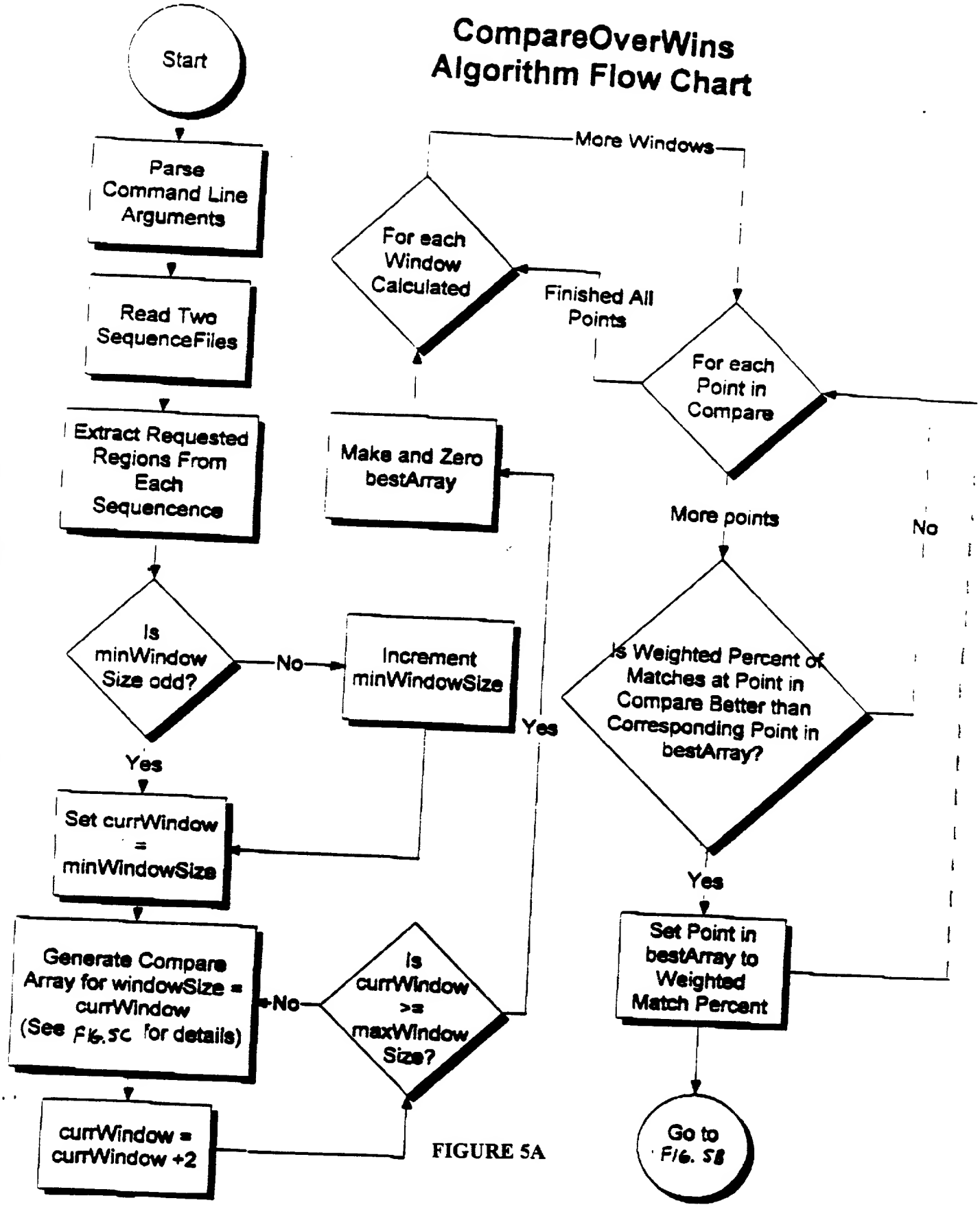


FIGURE 5A

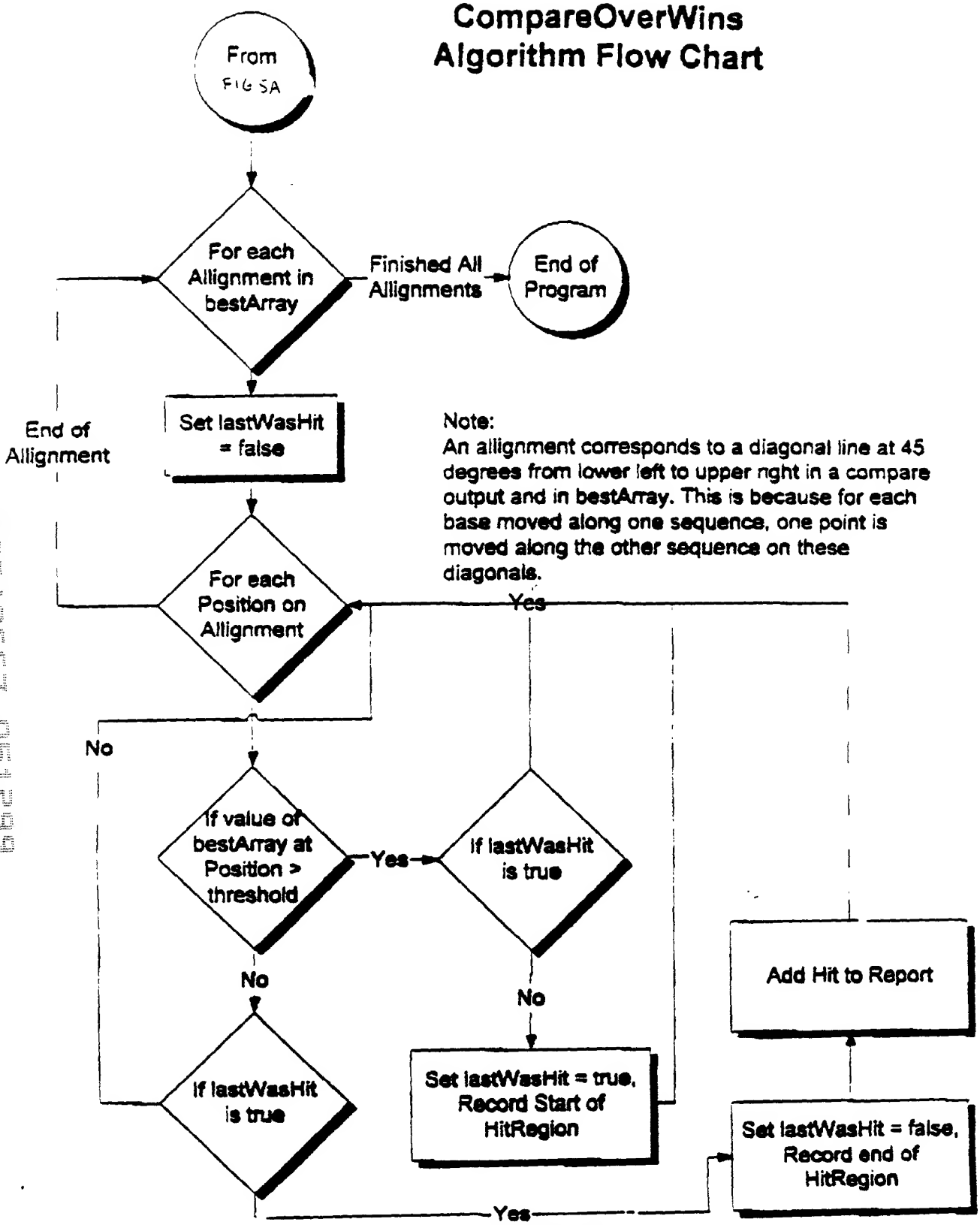

CompareOverWins
Algorithm Flow Chart
From FIG 5A
For each Allignment in bestArray
Finished All Allignments
End of Program
End of Allignment
Set lastWasHit = false
Note:
An allignment corresponds to a diagonal line at 45 degrees from lower left to upper right in a compare output and in bestArray. This is because for each base moved along one sequence, one point is moved along the other sequence on these diagonals.
For each Position on Allignment
Yes
No
If value of bestArray at Position > threshold
Yes
If lastWasHit is true
No
No
If lastWasHit is true
Set lastWasHit = true, Record Start of HitRegion
Add Hit to Report
Set lastWasHit = false, Record end of HitRegion
Yes

FIGURE 5B

CompareOverWins Algorithm Flow Chart Basic Compare

Input:
Sequence A length a
Sequence B length b
Window Size

Output:
Array of size a by b of unsigned chars (0-255)
Each point represents the number of matches in the window at that allignment and position

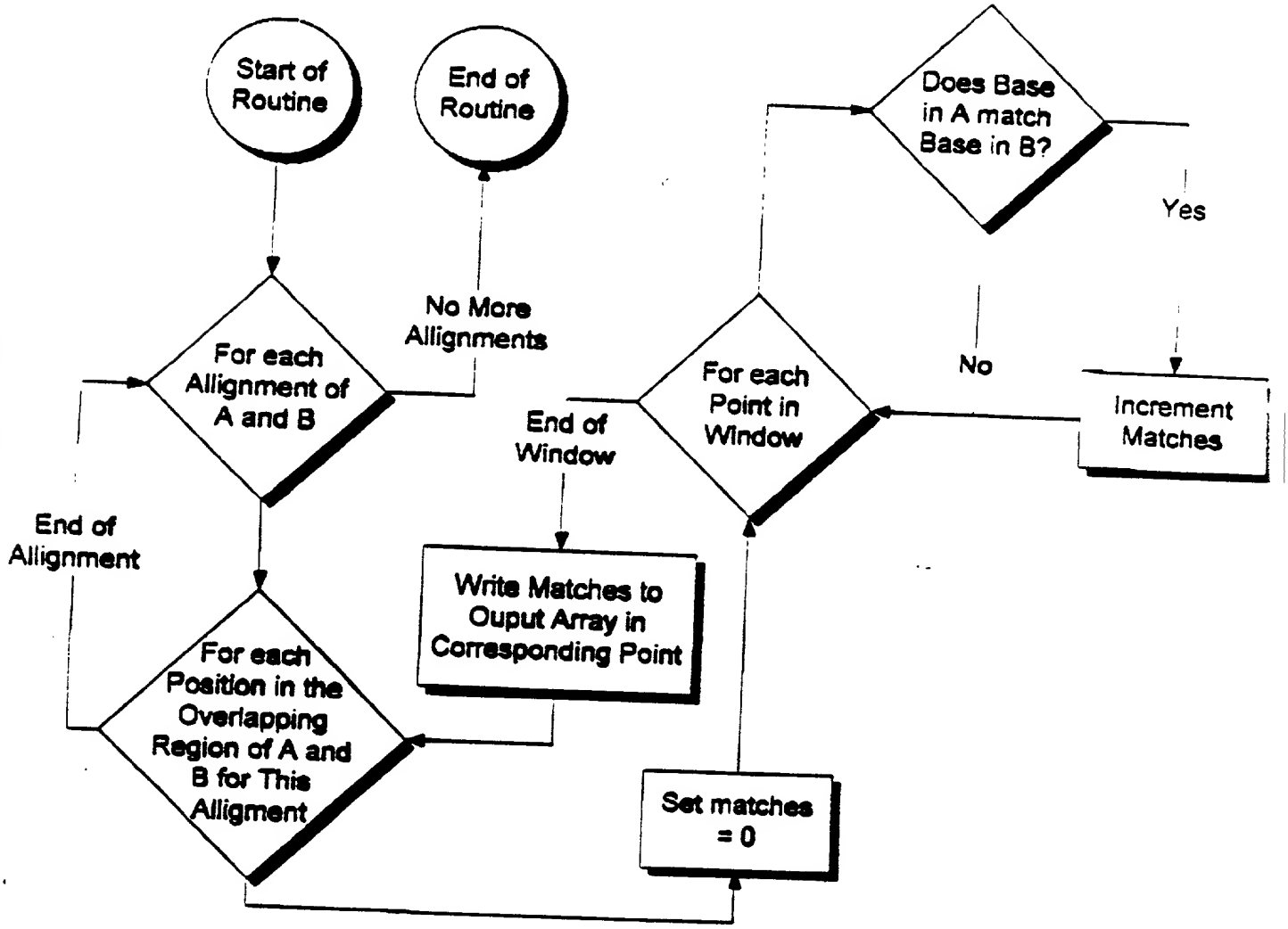
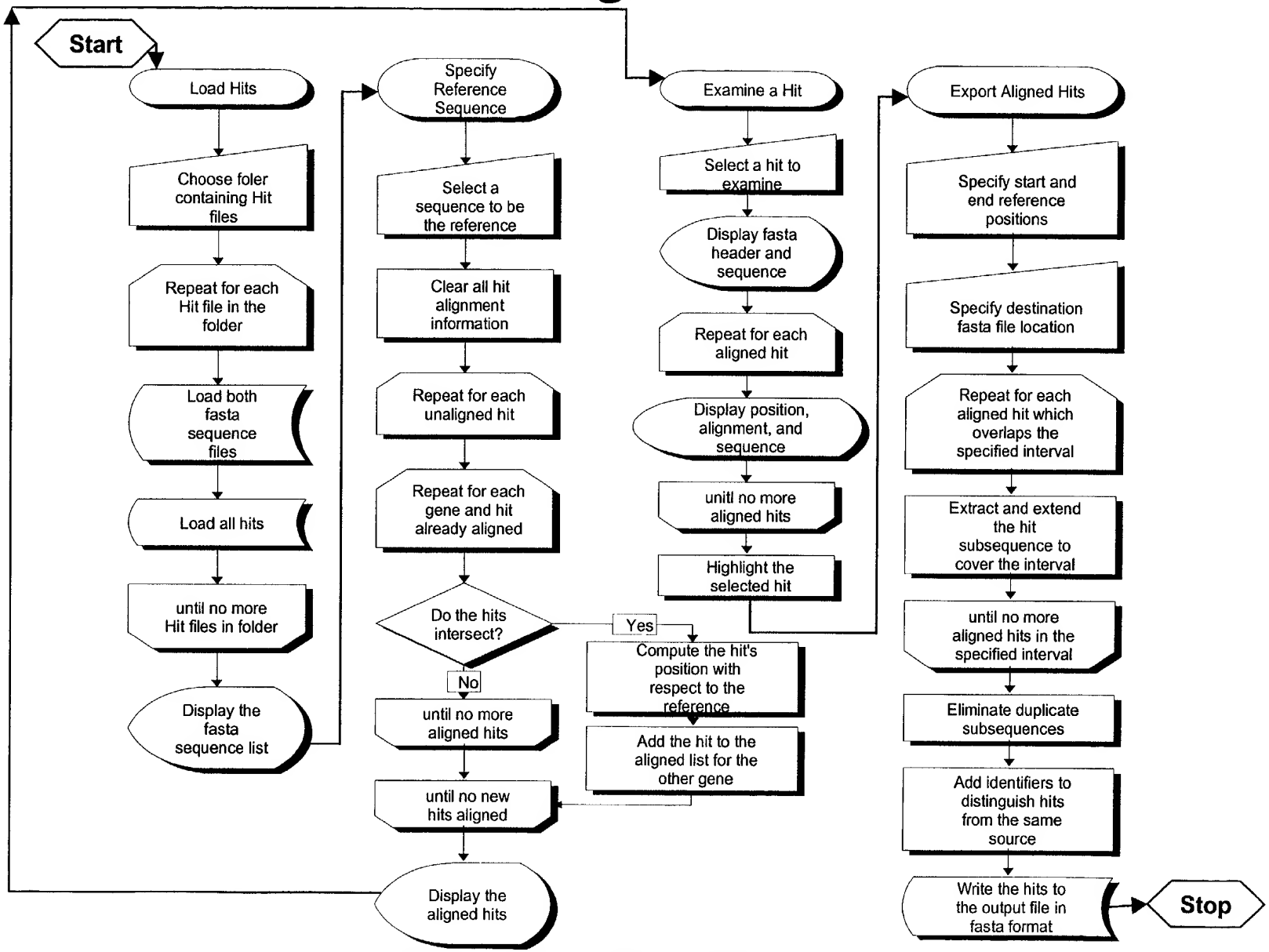


FIGURE 5C

Fig. 5D



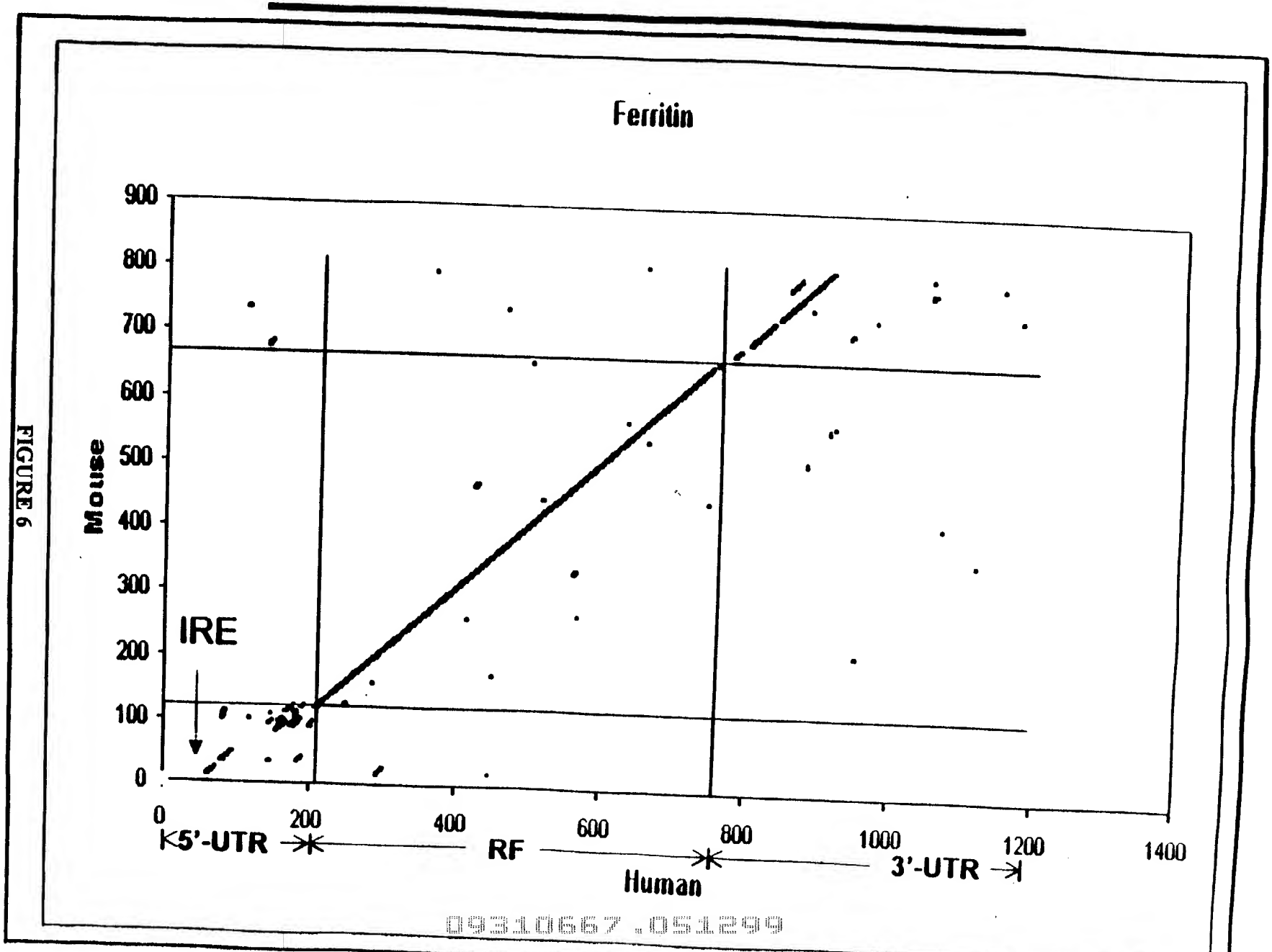


FIGURE 6

Self Complementation Analysis - Single Sequence

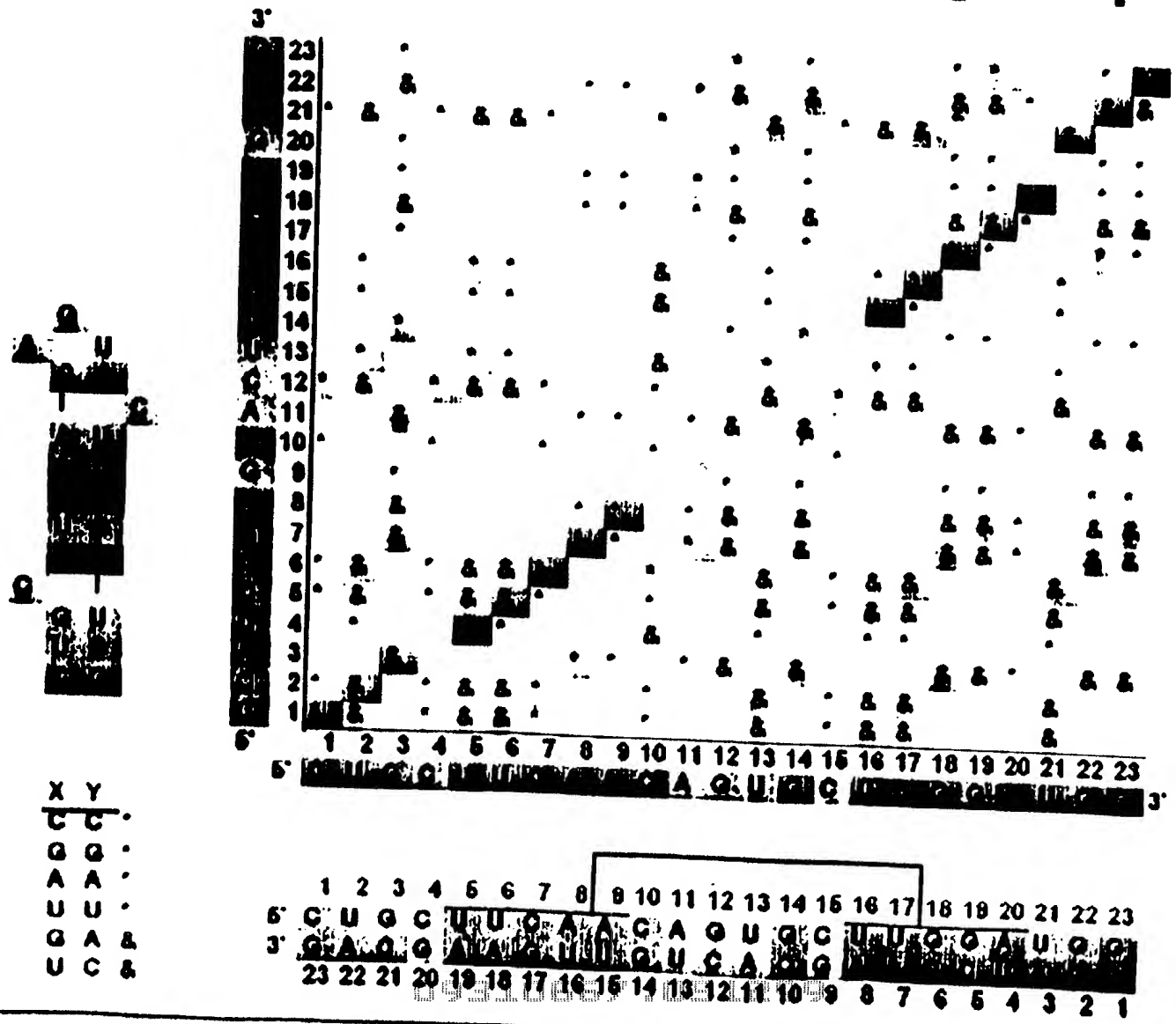


FIGURE 7

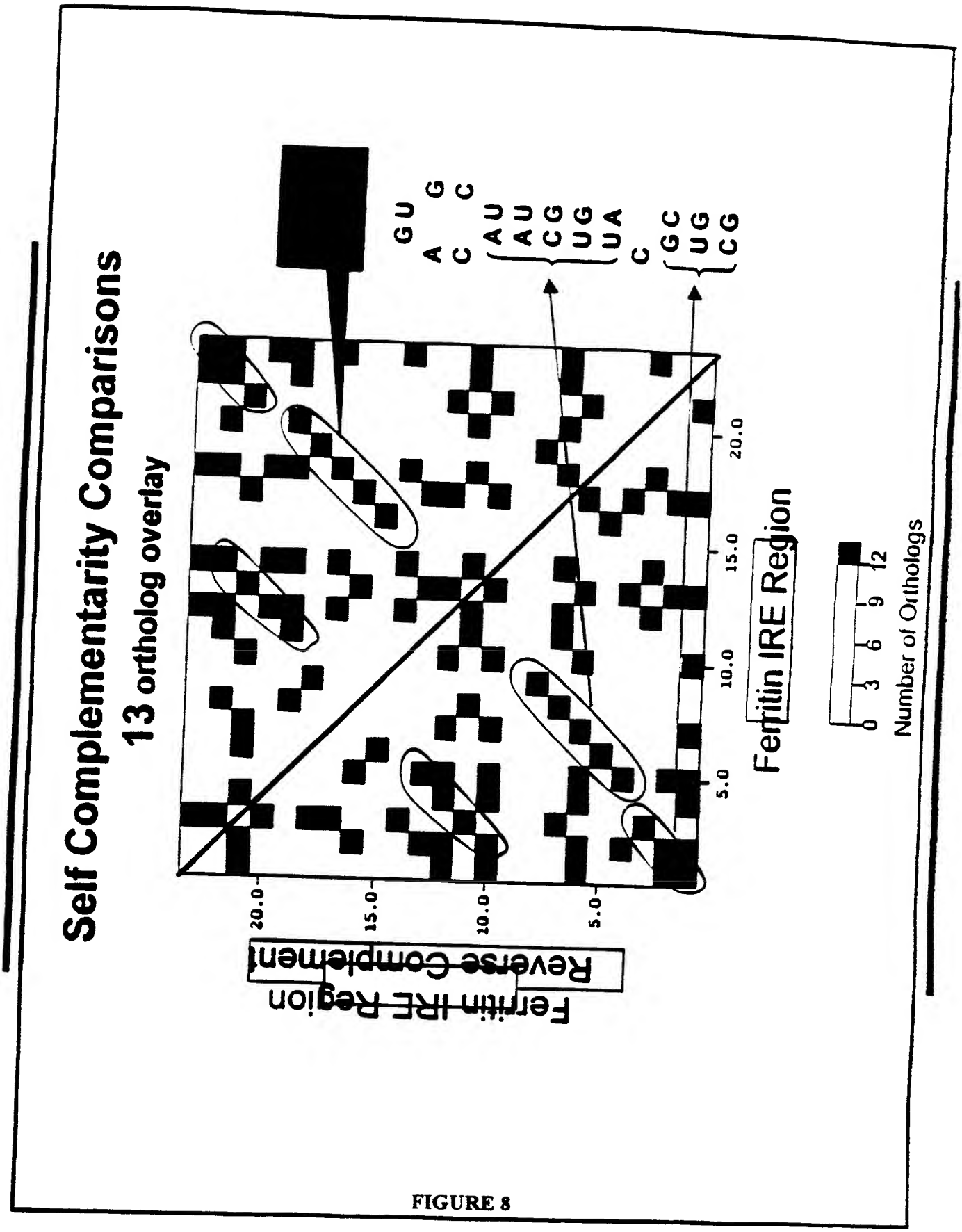
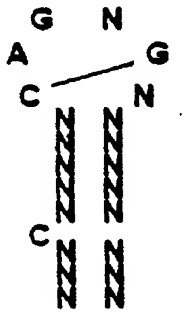

Self Complementarity Comparisons
13 ortholog overlay
Ferritin IRE Region
Reverse Complement
Ferritin IRE Region
Number of Orthologs
0 3 6 9 12
5.0
10.0
15.0
20.0
G U
A G
C C
A U
A U
C G
U G
U A
C
G C
U G
C G

FIGURE 8

Typical Descriptor

This is an example of a descriptor used to identify iron response elements.

To search the database using RNAMOT, the stem-loop model is converted to a text string as shown below:



IRE

Stem-loop Model

H1 S1 H2 S2 H2 H1

H1 3:3 NNN:NNN
S1 1 C
H2 5:5 NNNNN:NNNNN
S2 6 CAGNGN

W2
M0

IRE String descriptor

This descriptor allows for a wobble (W) of 2 (allows G-U pairing) and no mismatches. N can be any nucleotide
H refers to the stem region while S refers to the single stranded region.

FIGURE 9

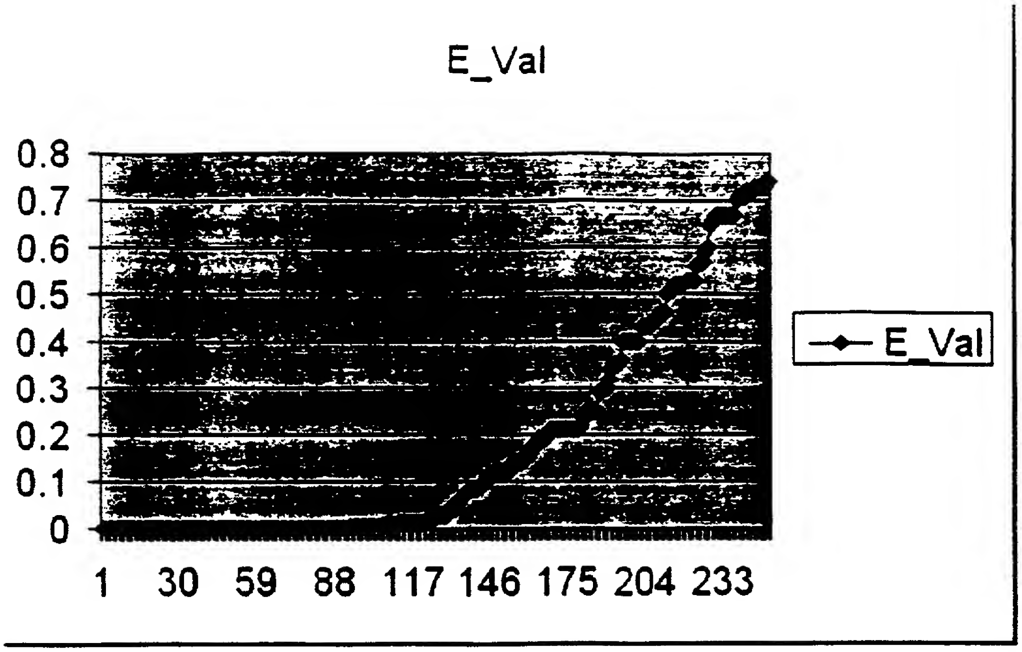
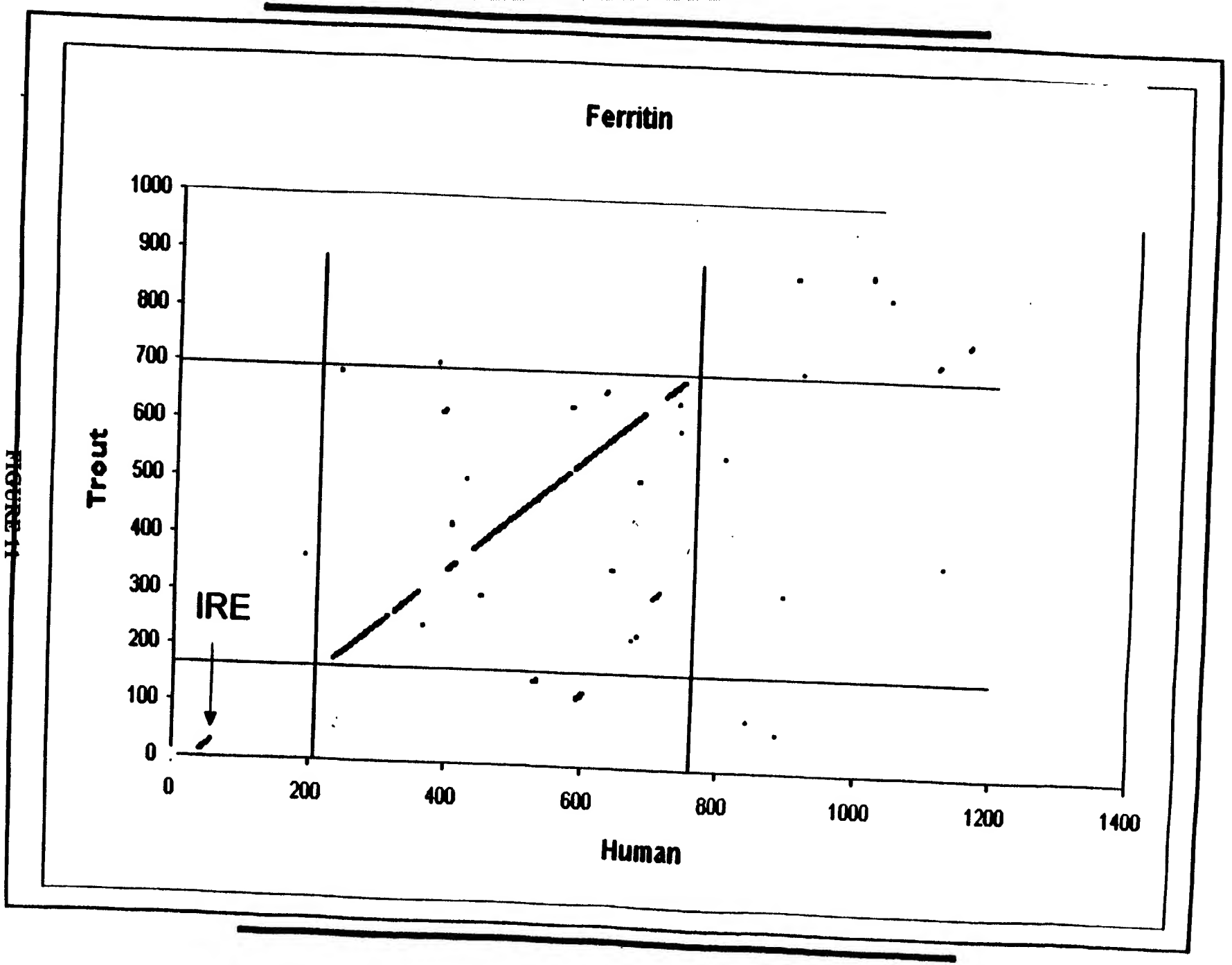


FIGURE 10



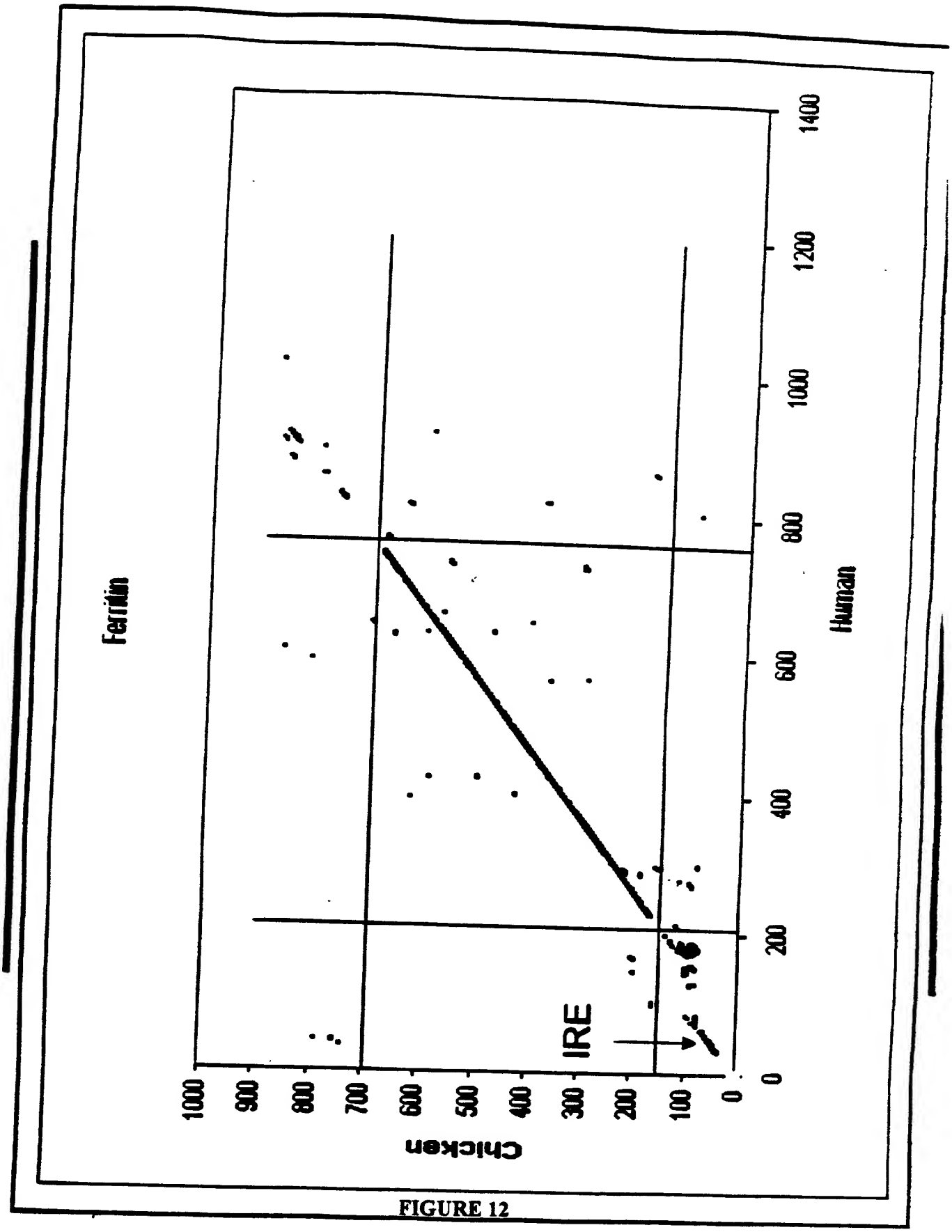
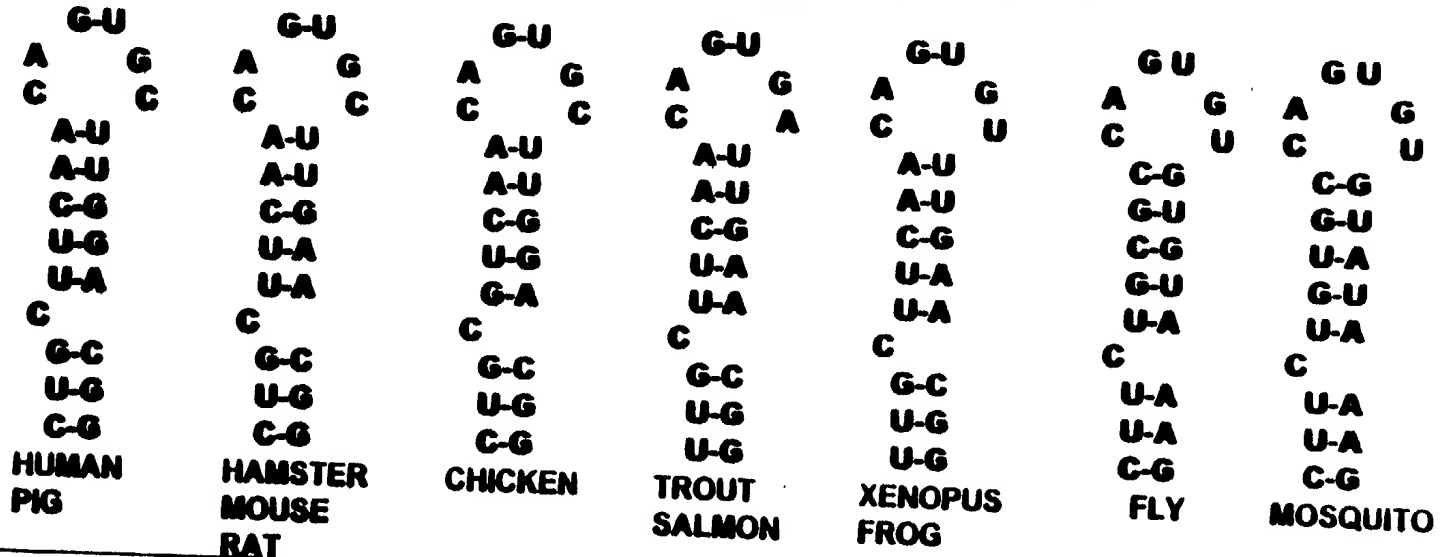

Ferritin
Chicken
Human
IRE
0
100
200
300
400
500
600
700
800
900
1000
0
200
400
600
800
1000
1200
1400

FIGURE 12



	HUMAN PIG	HAMSTER MOUSE RAT	CHICKEN	TROUT SALMON	XENOPUS FROG	FLY	MOSQUITO
HUMAN PIG	No	No	Yes	Yes	Yes	No	No
HAMSTER MOUSE RAT		No	Yes	Yes	Yes	No	No
CHICKEN			No	Yes	Yes	No	No
TROUT SALMON				No	Yes	Yes	Yes
XENOPUS FROG					No	Yes	Yes
FLY						No	Yes
MOSQUITO							No

FIGURE 13

Figure 14

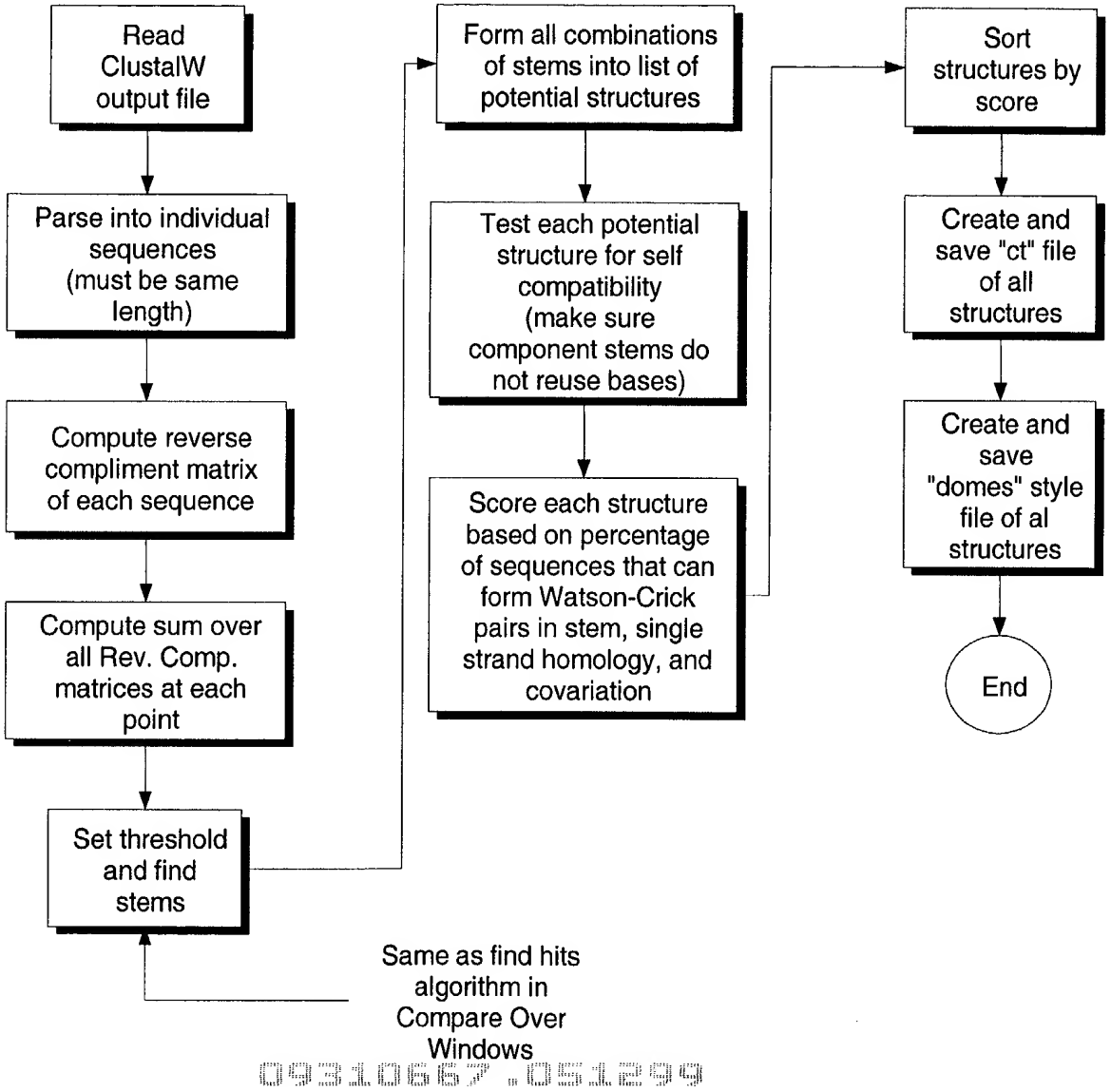


Figure 15

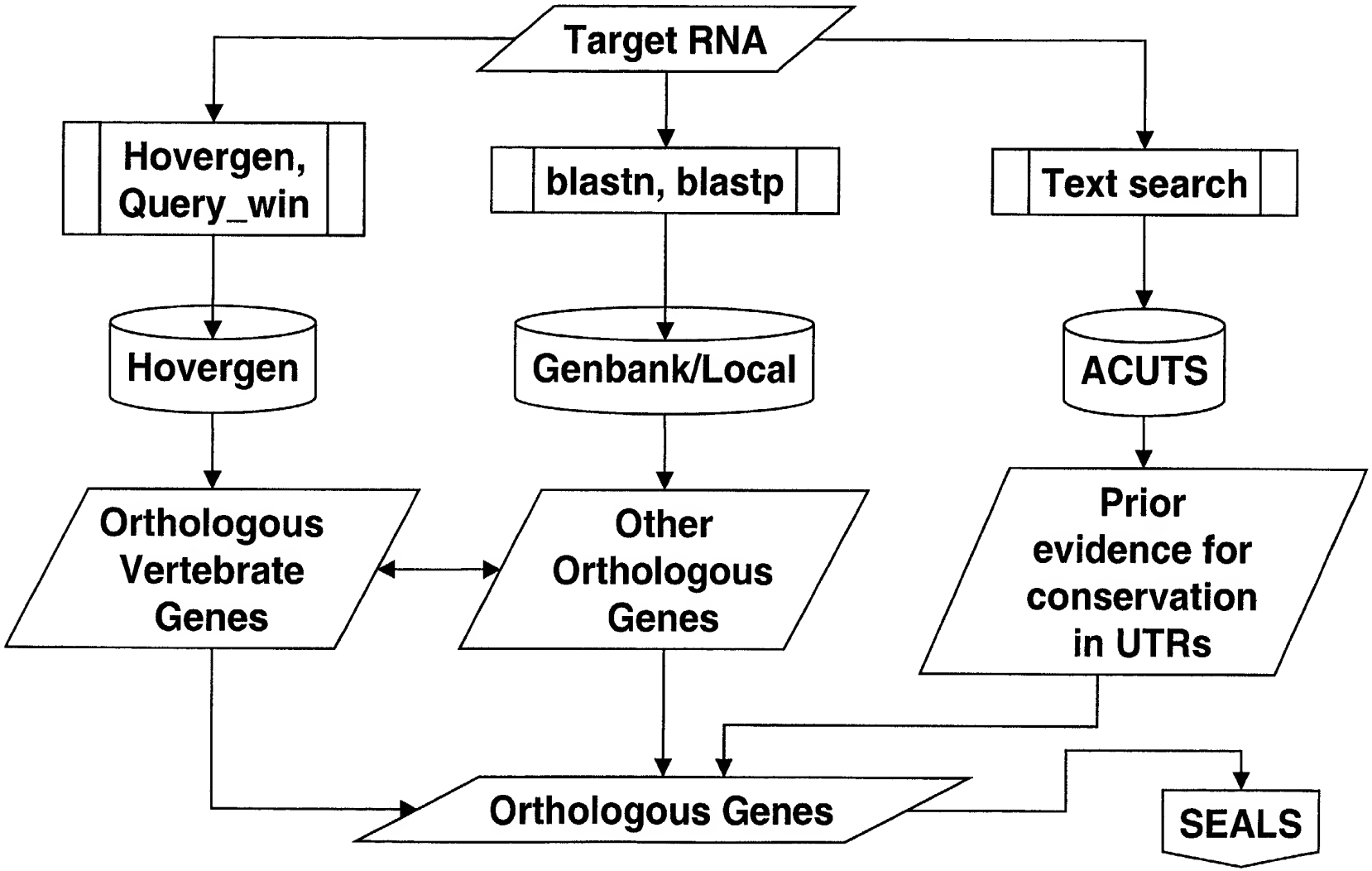


Figure 16

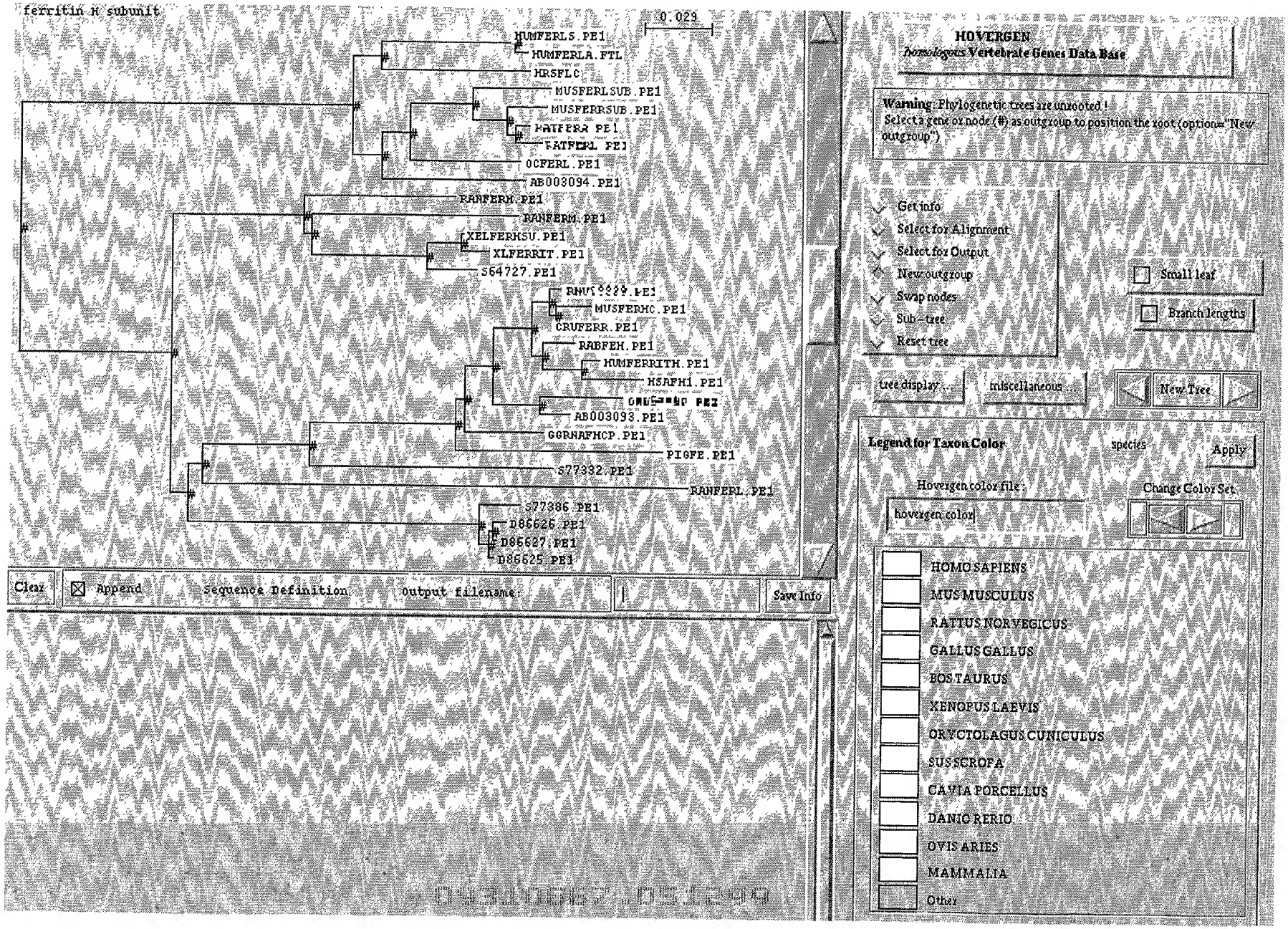

ferritin H subunit
0.029
HUMFERLS.PE1
HUMFERLA.FTL
MRSFLC
MUSFERLSUB.PE1
MUSFERRSUB.PE1
RATFERR.PE1
RATFERL.PE1
OCFERL.PE1
AB003094.PE1
RANFERH.PE1
RANFERM.PE1
XELFERHSU.PE1
XLFERRIT.PE1
S64727.PE1
MUSFERHC.PE1
CRUFERR.PE1
RABFEH.PE1
HUMFERRITH.PE1
HSAFH1.PE1
AB003093.PE1
GGRNAFHCP.PE1
PIGFE.PE1
S77332.PE1
RANFERL.PE1
S77386.PE1
D86626.PE1
D86627.PE1
D86625.PE1
Clear
Append
Sequence Definition
Output filename :
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted !
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display
miscellaneous
New Tree
Legend for Taxon Color
species
Apply
Hovergen color file :
hovergen.color
Change Color Set
HOMO SAPIENS
MUS MUSCULUS
RATTUS NORVEGICUS
GALLUS GALLUS
BOS TAURUS
XENOPUS LAEVIS
ORYCTOLAGUS CUNICULUS
SUS SCROFA
CAVIA PORCELLUS
DANIO RERIO
OVIS ARIES
MAMMALIA
Other

Figure 17

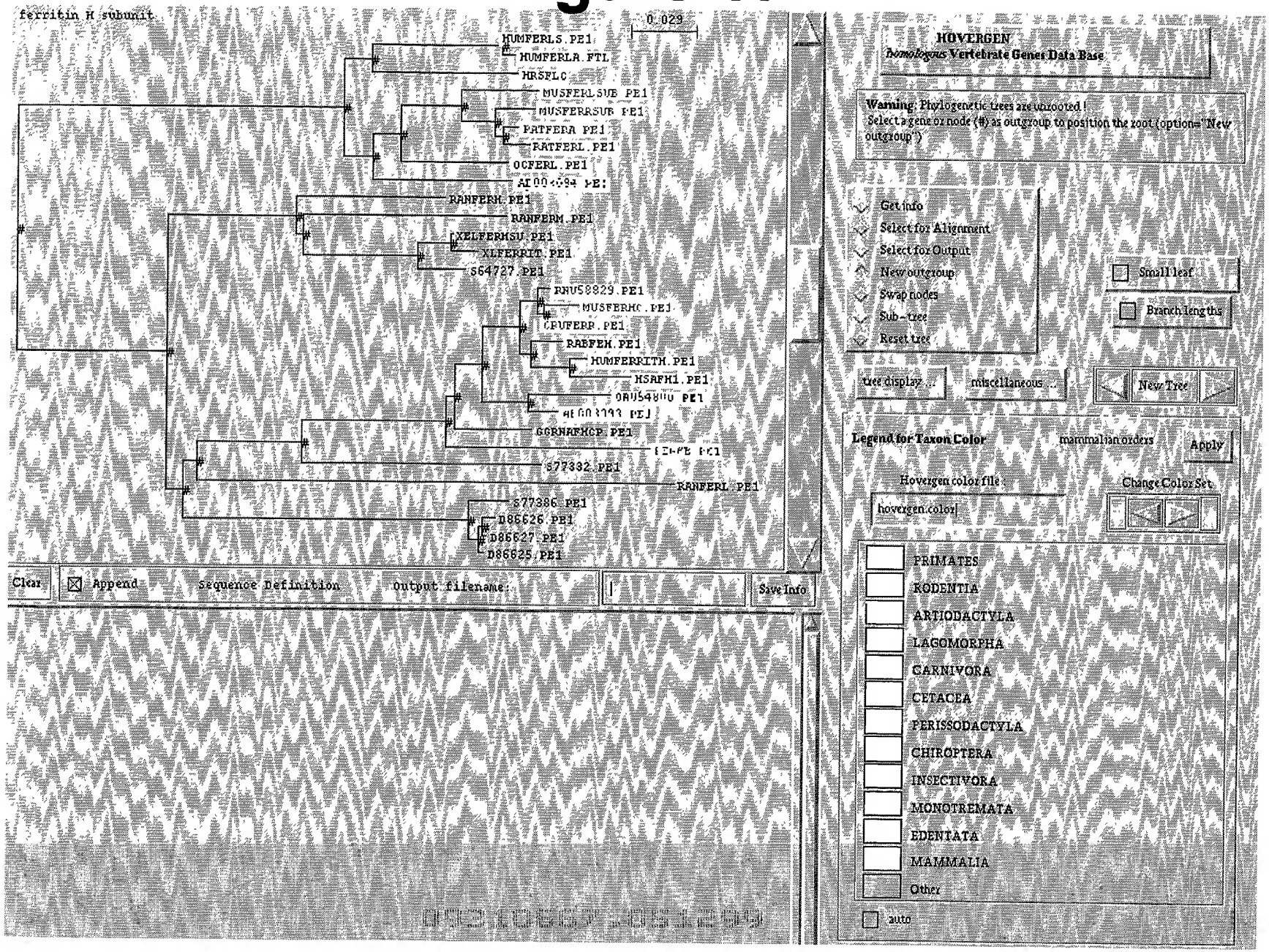

ferritin H subunit
0.029
HUMFERL5.PE1
HUMFERLA.FTL
MRSFLC
MUSFERL5UB.PE1
MUSFERRSUB.PE1
RATFERA.PE1
RATFERL.PE1
OCFERL.PE1
RANFERH.PE1
RANFERM.PE1
XELFERHSU.PE1
XLFERRIT.PE1
S64727.PE1
RNU58829.PE1
MUSFERHC.PE1
CPUFERP.PE1
RABFERH.PE1
HUMFERRITH.PE1
MSAFH1.PE1
GGRNAFHCP.PE1
S77332.PE1
RANFERL.PE1
S77386.PE1
D86626.PE1
D86627.PE1
D86625.PE1
Clear
Append
Sequence Definition
Output filename:
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted!
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display ...
miscellaneous ...
New Tree
Legend for Taxon Color
mammalian orders
Apply
Hovergen color file
Change Color Set
hovergen.color
PRIMATES
RODENTIA
ARTIODACTYLA
LAGOMORPHA
CARNIVORA
CETACEA
PERISSODACTYLA
CHIROPTERA
INSECTIVORA
MONOTREMATA
EDENTATA
MAMMALIA
Other
auto

Figure 18

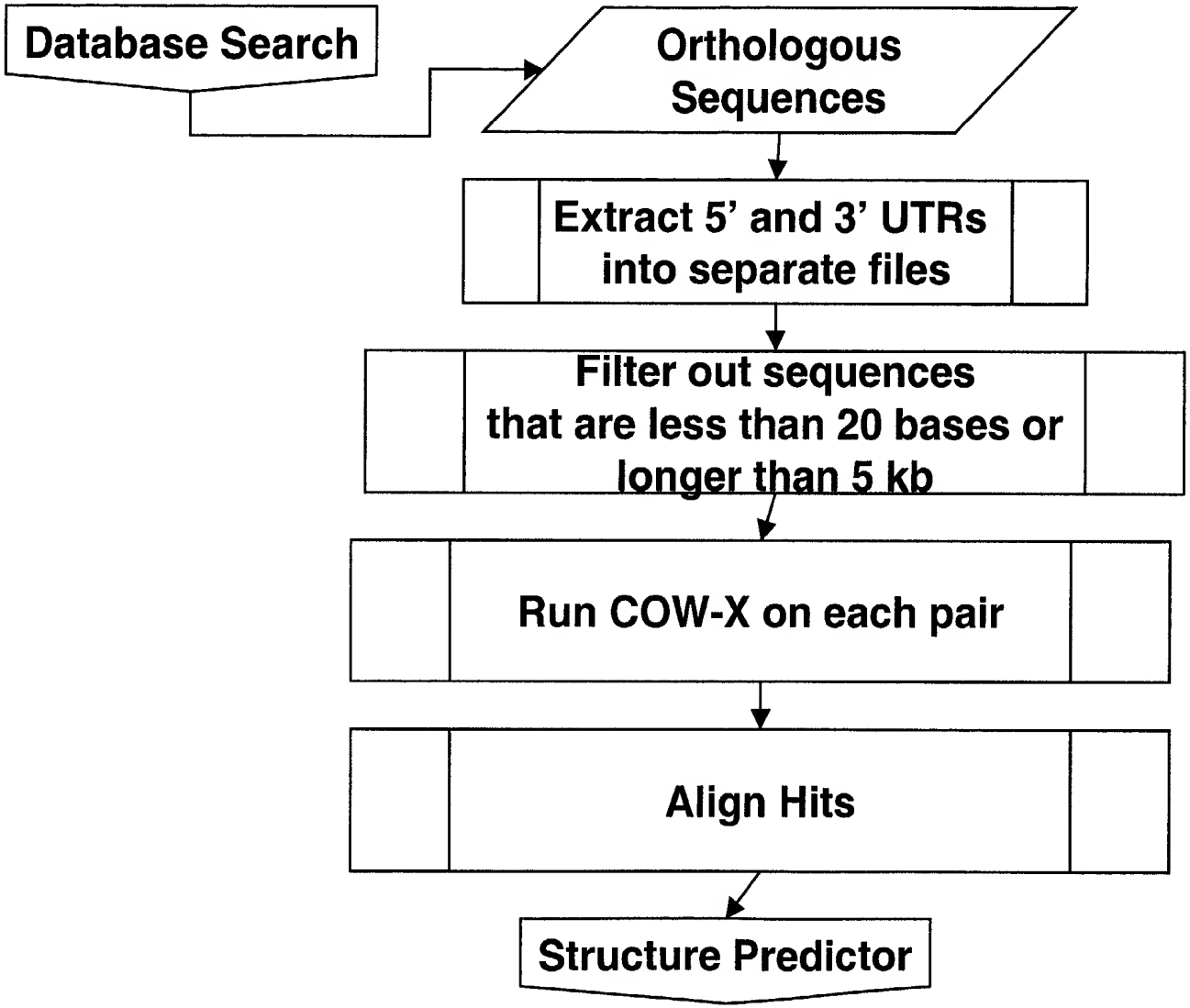


Figure 19

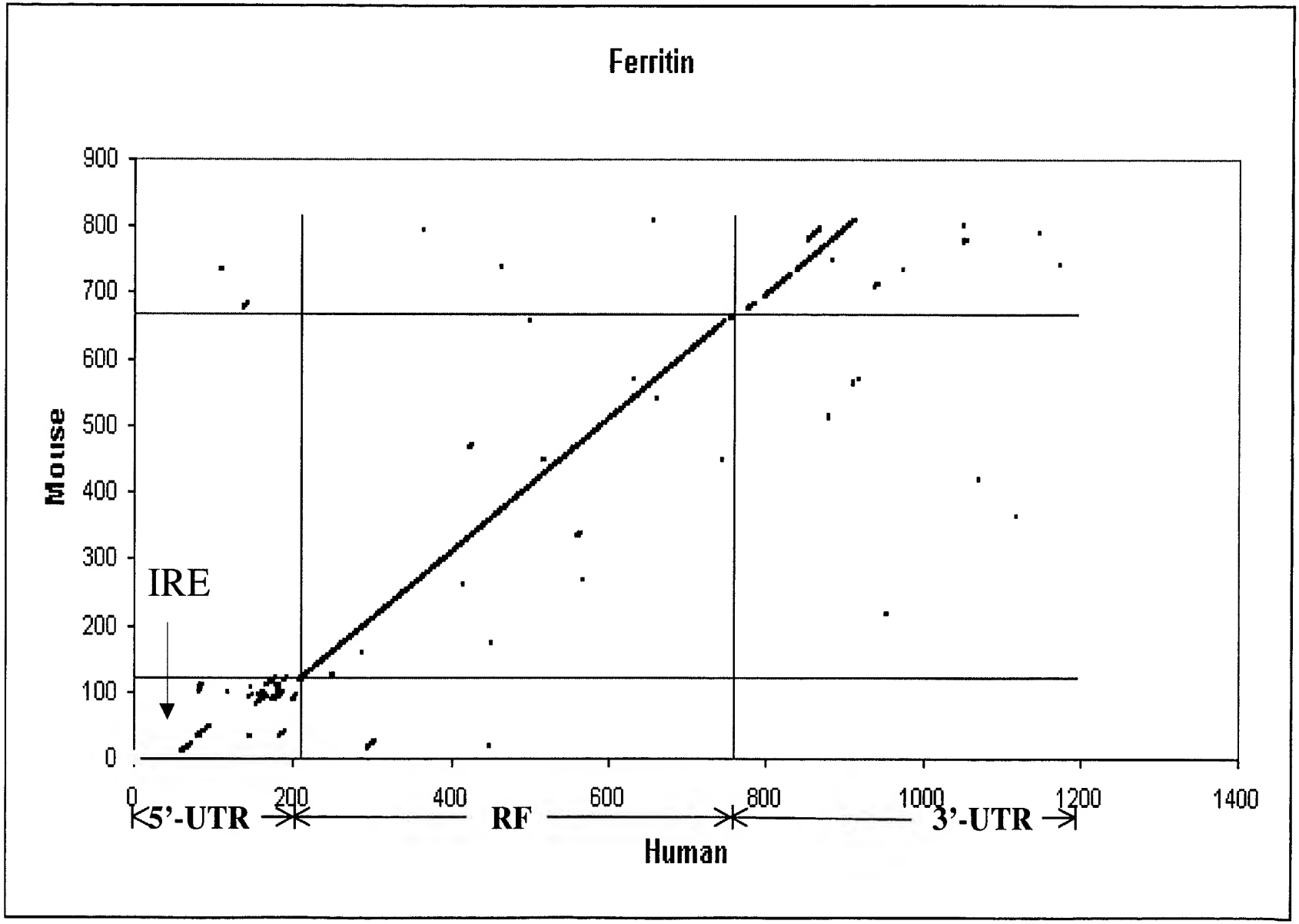


Figure 20

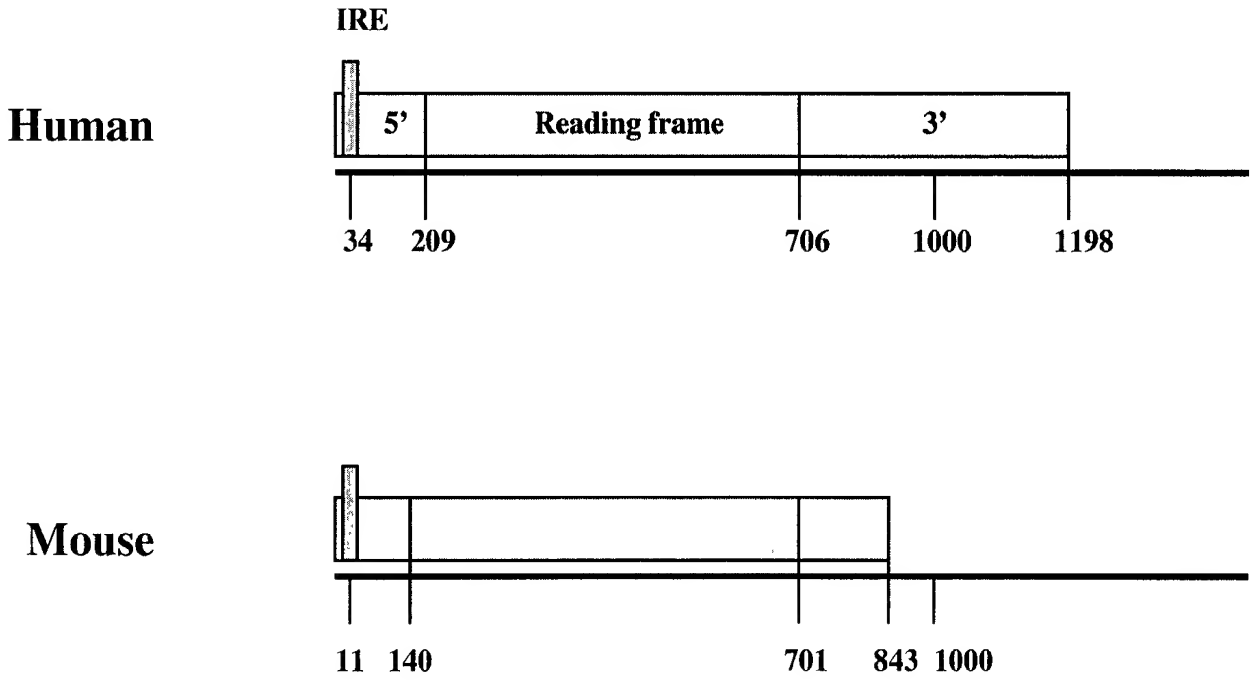


Figure 21

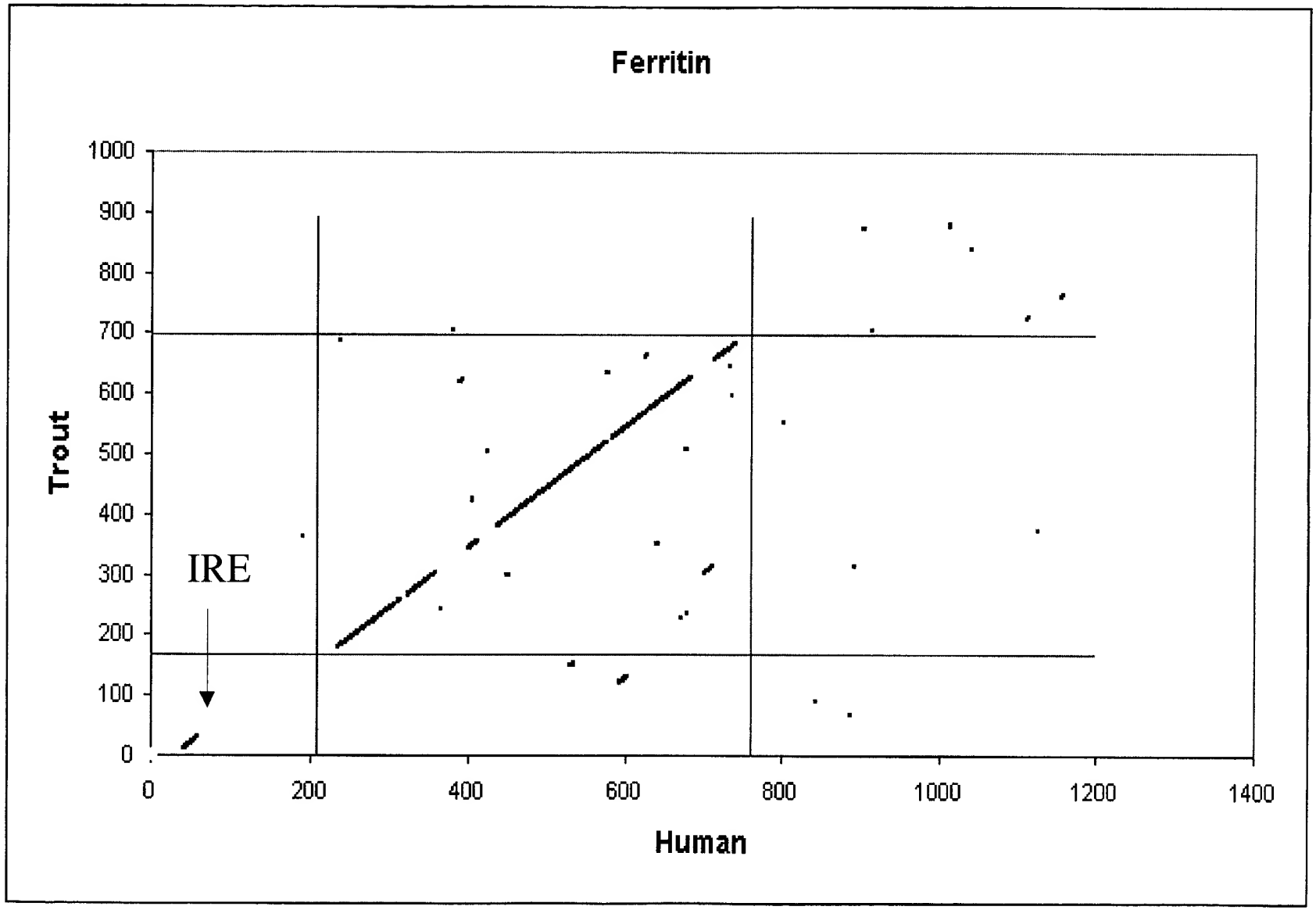


Figure 22

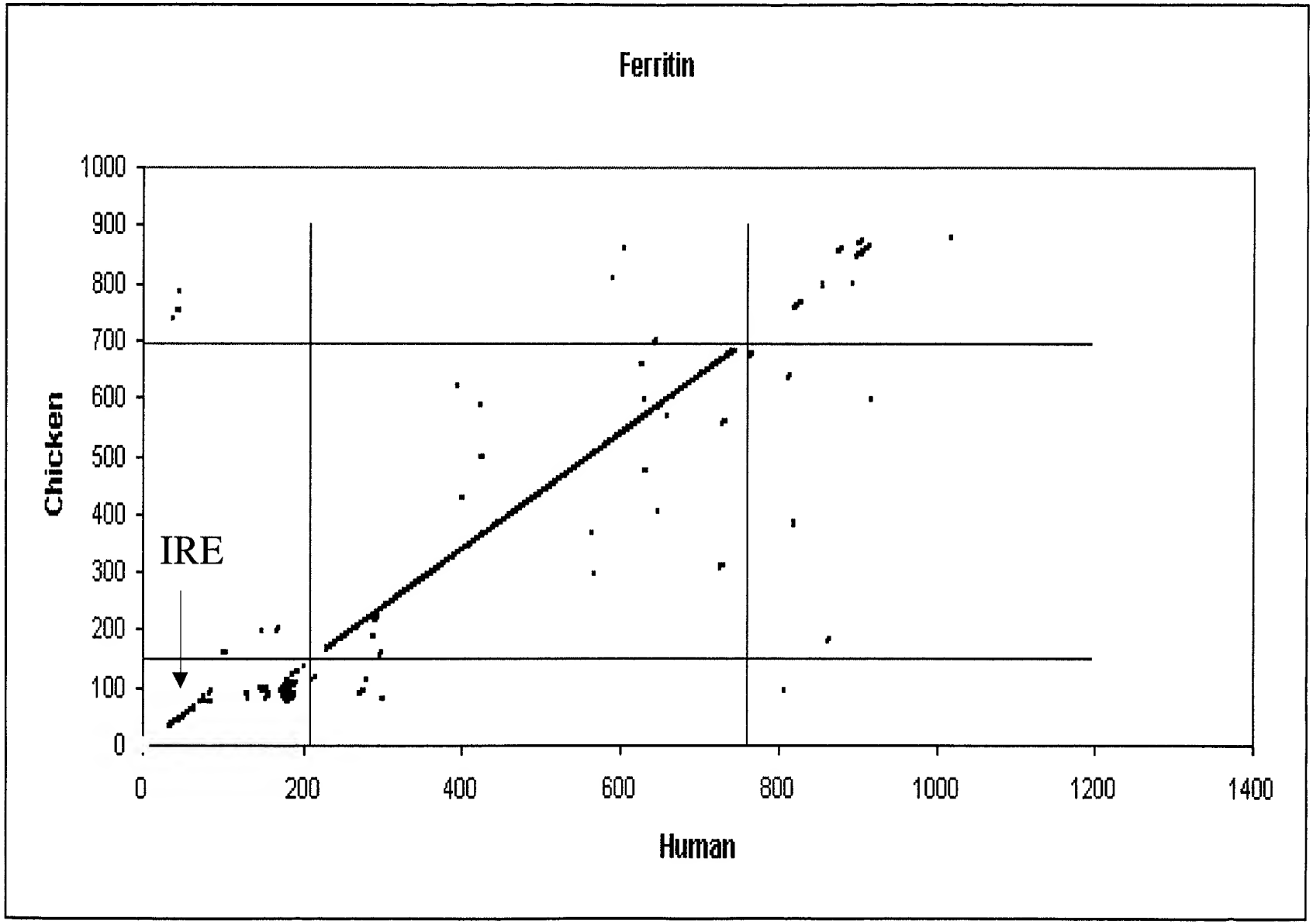


Figure 23

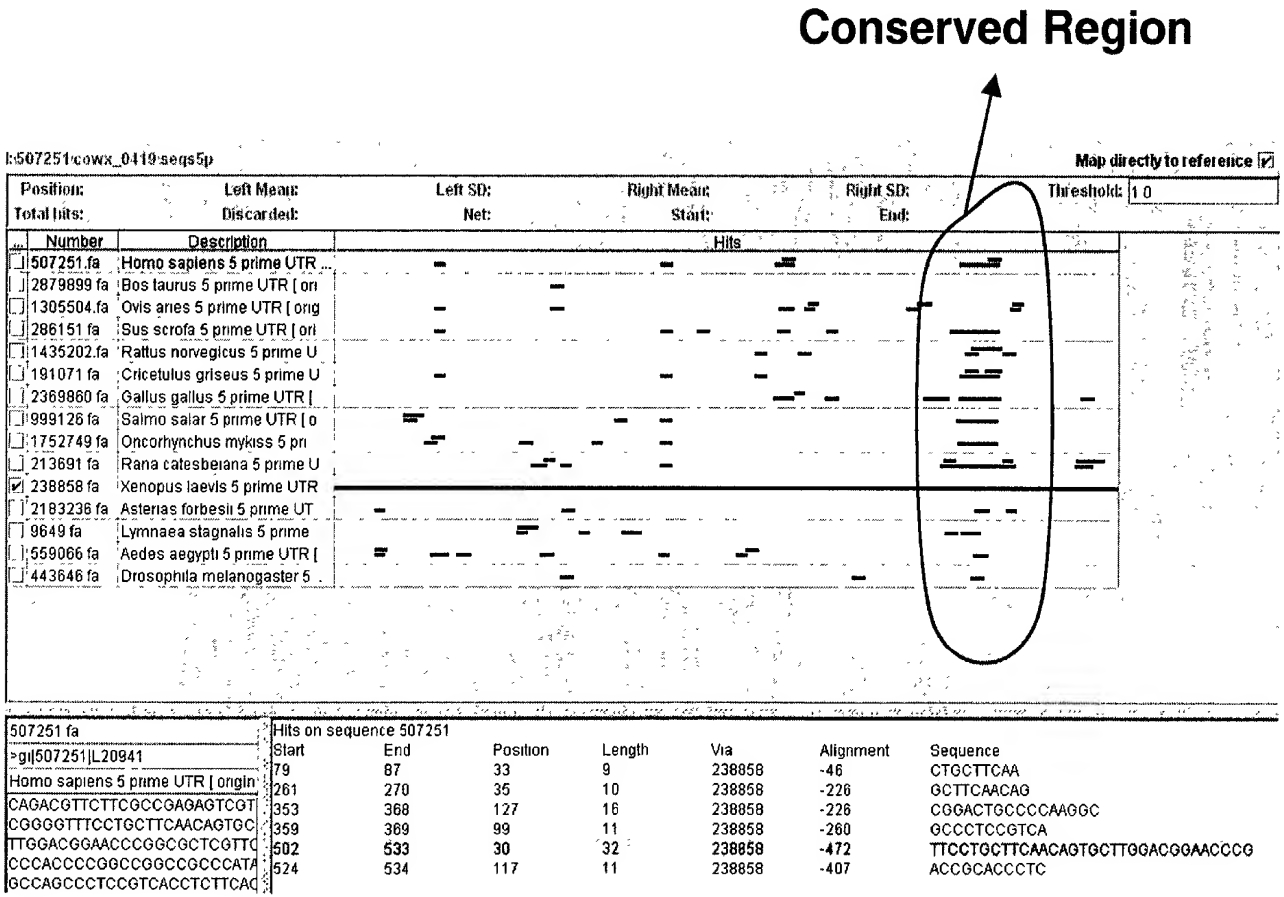
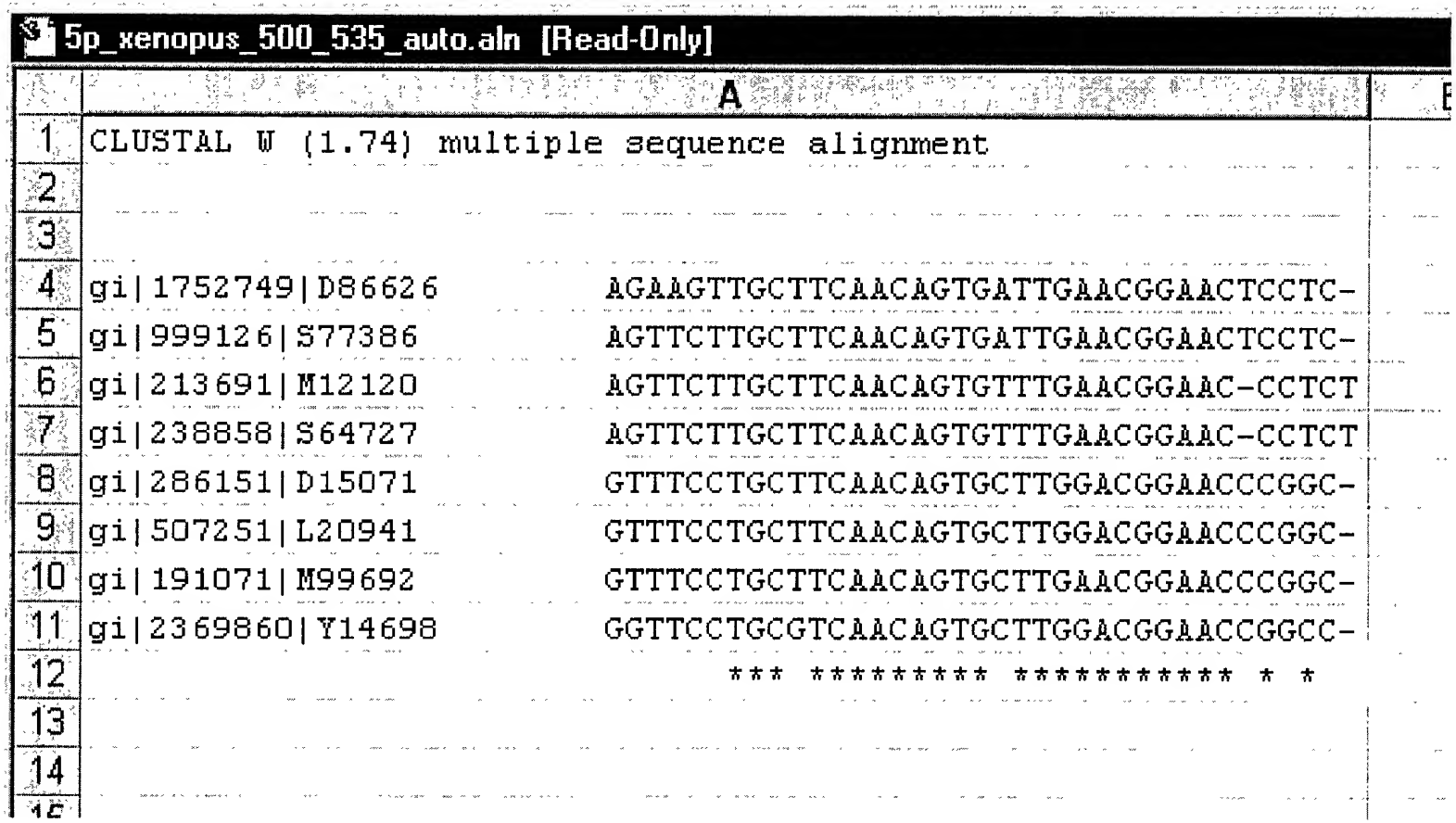

Conserved Region
Map directly to reference
Position:
Left Mean:
Left SD:
Right Mean:
Right SD:
Threshold: 1.0
Total hits:
Discarded:
Net:
Start:
End:
Number
Description
Hits
507251.fa Homo sapiens 5 prime UTR ...
2879899.fa Bos taurus 5 prime UTR [ori
1305504.fa Ovis aries 5 prime UTR [orig
286151.fa Sus scrofa 5 prime UTR [ori
1435202.fa Rattus norvegicus 5 prime U
191071.fa Cricetulus griseus 5 prime U
2369860.fa Gallus gallus 5 prime UTR [
999126.fa Salmo salar 5 prime UTR [o
1752749.fa Oncorhynchus mykiss 5 pri
213691.fa Rana catesbeiana 5 prime U
238858.fa Xenopus laevis 5 prime UTR
2183236.fa Asterias forbesii 5 prime UT
9649.fa Lymnaea stagnalis 5 prime
559066.fa Aedes aegypti 5 prime UTR [
443646.fa Drosophila melanogaster 5 .
507251.fa
>gi|507251|L20941
Homo sapiens 5 prime UTR [origin
CAGACGTTCTTCGCCGAGAGTCGT
CGGGGTTTCCTGCTTCAACAGTGC
TTGGACGGAACCCGGCGCTCGTTC
CCCACCCCGGCCGGCCGCCCATA
GCCAGCCCTCCGTCACCTCTTCAC
Hits on sequence 507251
Start End Position Length Via Alignment Sequence
79 87 33 9 238858 -46 CTGCTTCAA
261 270 35 10 238858 -226 GCTTCAACAG
353 368 127 16 238858 -226 CGGACTGCCCCAAGGC
359 369 99 11 238858 -260 GCCCTCCGTCA
502 533 30 32 238858 -472 TTCCTGCTTCAACAGTGCTTGGACGGAACCCG
524 534 117 11 238858 -407 ACCGCACCCTC

Figure 24



5p_xenopus_500_535_auto.aln [Read-Only]

	A
1	CLUSTAL W (1.74) multiple sequence alignment
2	
3	
4	gi|1752749|D86626 AGAAGTTGCTTCAACAGTGATTGAACGGAACTCCTC-
5	gi|999126|S77386 AGTTCTTGCTTCAACAGTGATTGAACGGAACTCCTC-
6	gi|213691|M12120 AGTTCTTGCTTCAACAGTGTTTGAACGGAAC-CCTCT
7	gi|238858|S64727 AGTTCTTGCTTCAACAGTGTTTGAACGGAAC-CCTCT
8	gi|286151|D15071 GTTTCCTGCTTCAACAGTGCTTGGACGGAACCCGGC-
9	gi|507251|L20941 GTTTCCTGCTTCAACAGTGCTTGGACGGAACCCGGC-
10	gi|191071|M99692 GTTTCCTGCTTCAACAGTGCTTGAACGGAACCCGGC-
11	gi|2369860|Y14698 GGTTCCTGCGTCAACAGTGCTTGGACGGAACCGGCC-
12	*** ********* ********** * *
13	
14	

Figure 25

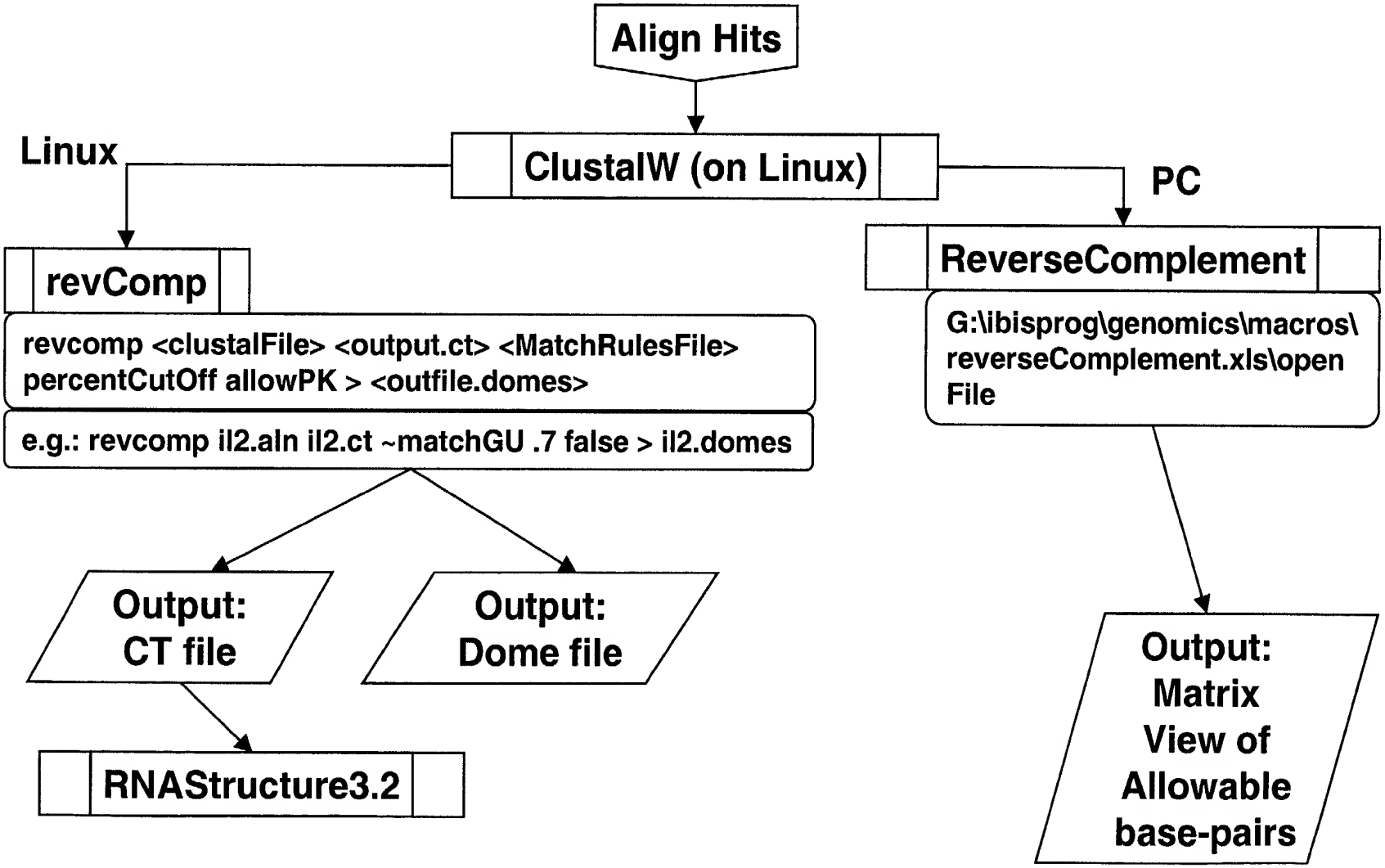


Figure 26

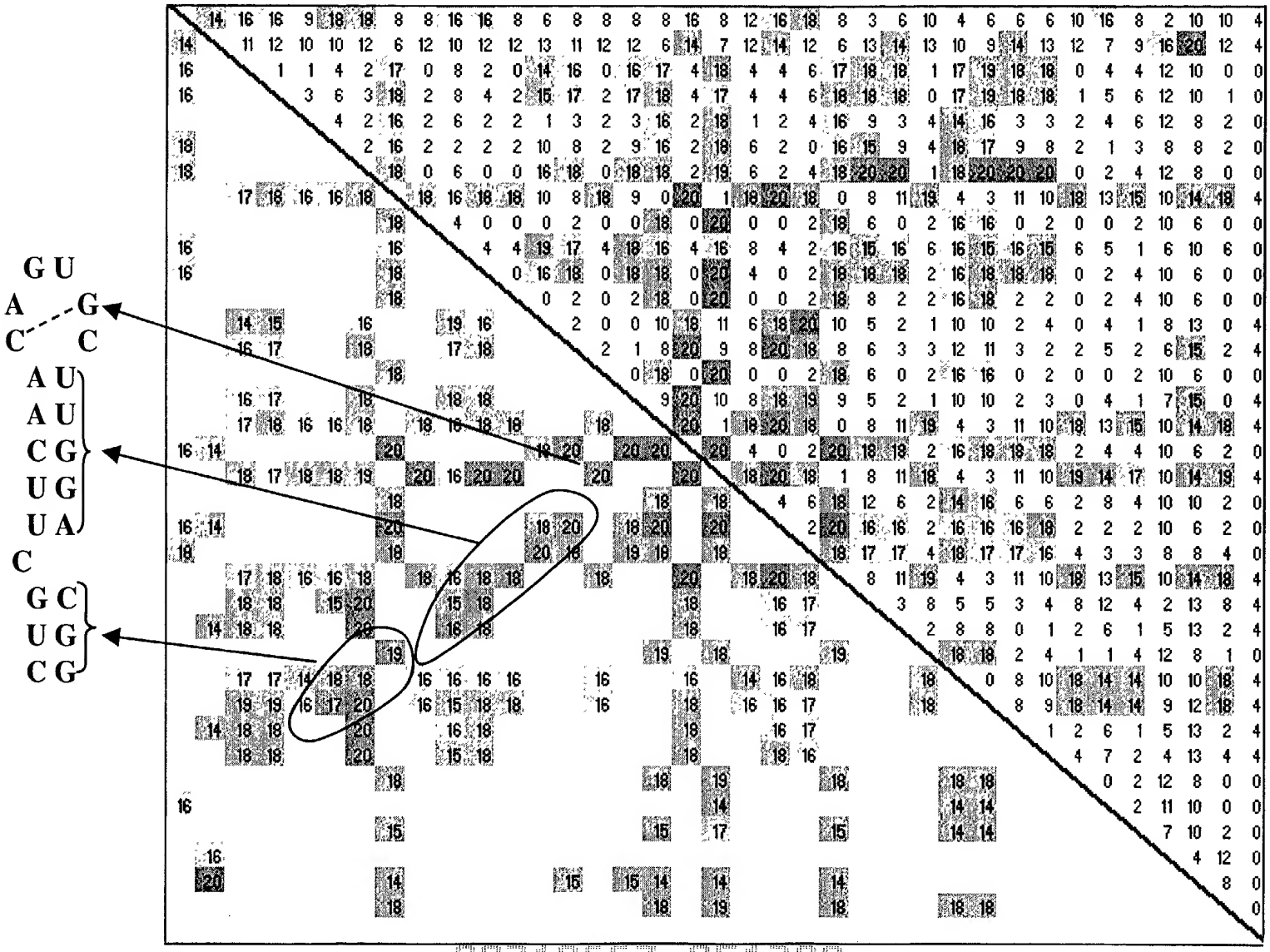


Figure 27

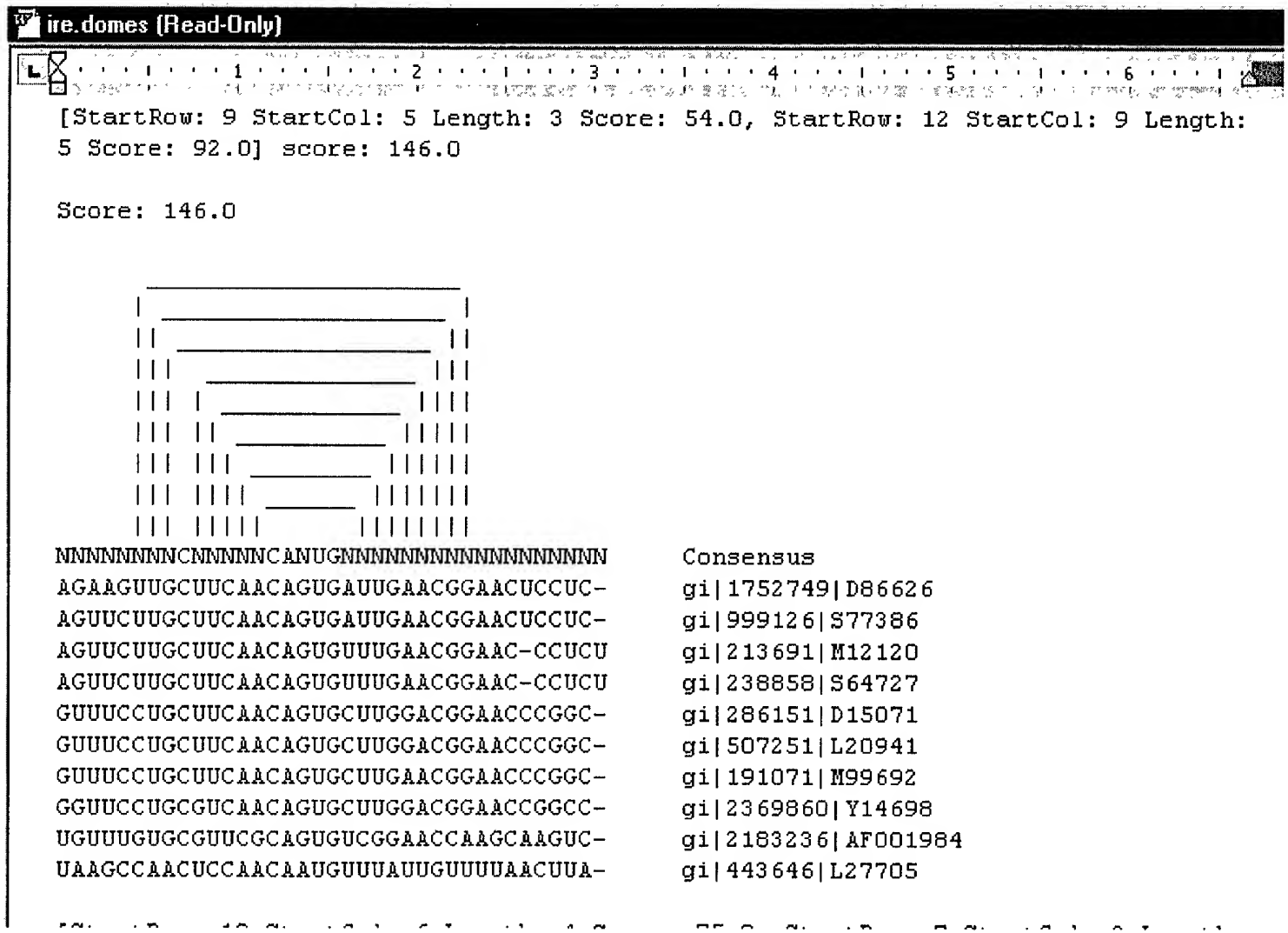


Figure 28

IRE Structures for each species

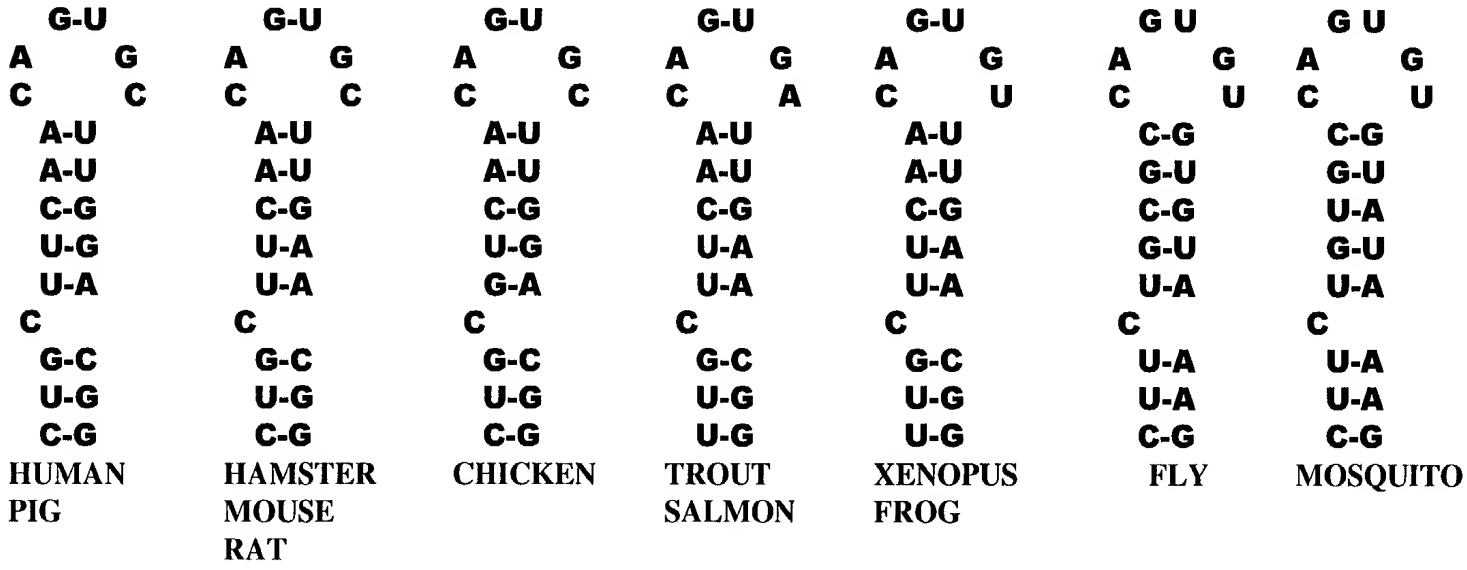


Figure 29

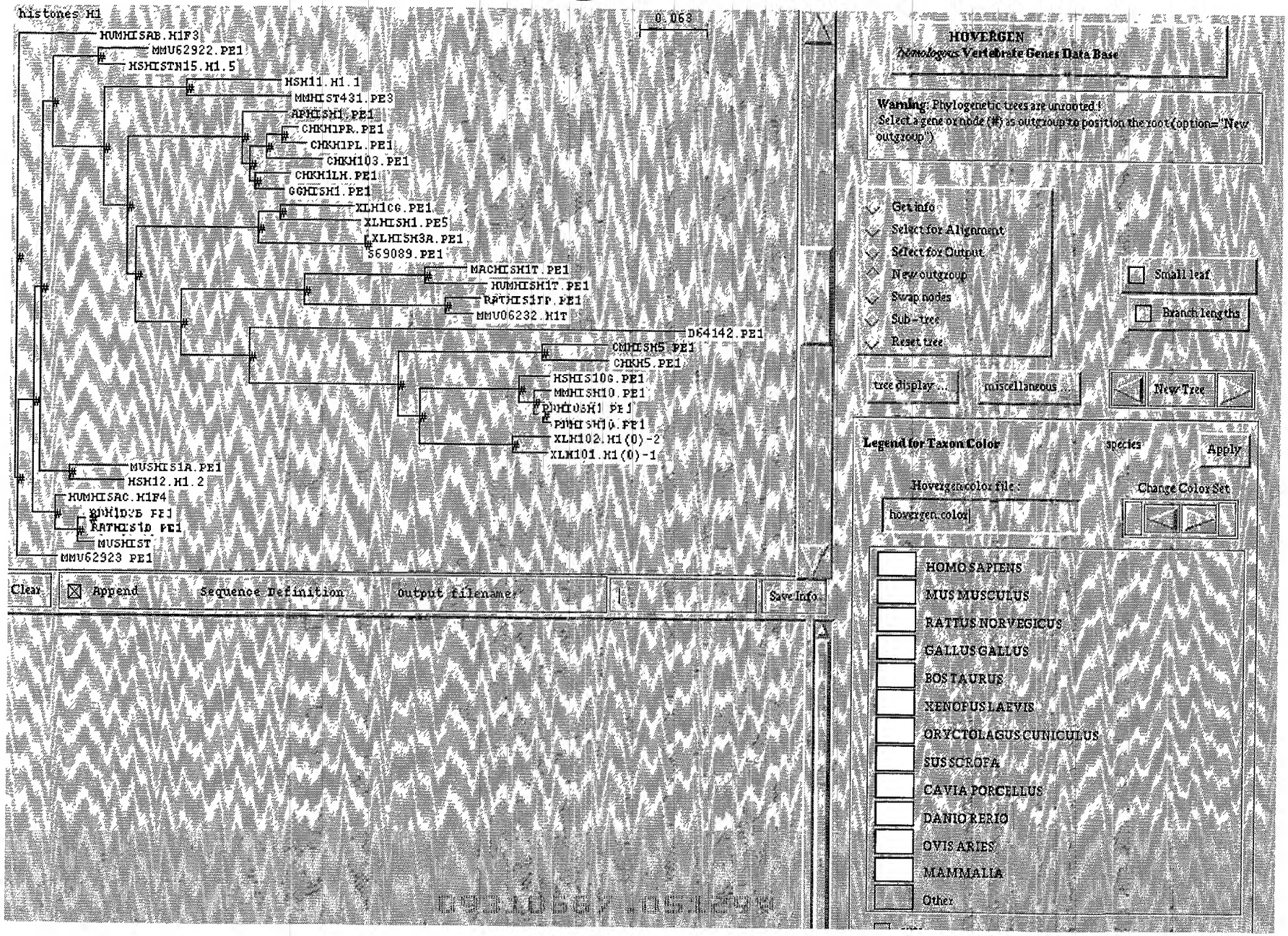

histones H1
0.068
HUMHISAB.H1F3
MMU62922.PE1
HSHISTN15.H1.5
HSH11.H1.1
MMHIST431.PE3
APHISH1.PE1
CHKH1PR.PE1
CHKH1PL.PE1
CHKH103.PE1
CHKH1LH.PE1
GGHISH1.PE1
XLH106.PE1
XLHISM1.PE5
XLHISM3A.PE1
S69089.PE1
MACHISH1T.PE1
HUMHISH1T.PE1
RATHIS1TP.PE1
MMU06232.H1T
D64142.PE1
CHKH5.PE1
HSHIS10G.PE1
MMHISH10.PE1
XLH102.H1(0)-2
XLH101.H1(0)-1
MUSHIS1A.PE1
HSH12.H1.2
HUMHISAC.H1F4
RATHIS1D.PE1
MUSHIST
MMU62923.PE1
Clear
Append
Sequence Definition
Output filename
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted!
Select a gene or node (#) as outgroup to position the root (option='New outgroup')
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display ...
miscellaneous ...
New Tree
Legend for Taxon Color
species
Apply
Hovergen color file:
hovergen.colors
Change Color Set
HOMO SAPIENS
MUS MUSCULUS
RATTUS NORVEGICUS
GALLUS GALLUS
BOS TAURUS
XENOPUS LAEVIS
ORYCTOLAGUS CUNICULUS
SUS SCROFA
CAVIA PORCELLUS
DANIO RERIO
OVIS ARIES
MAMMALIA
Other

Figure 30

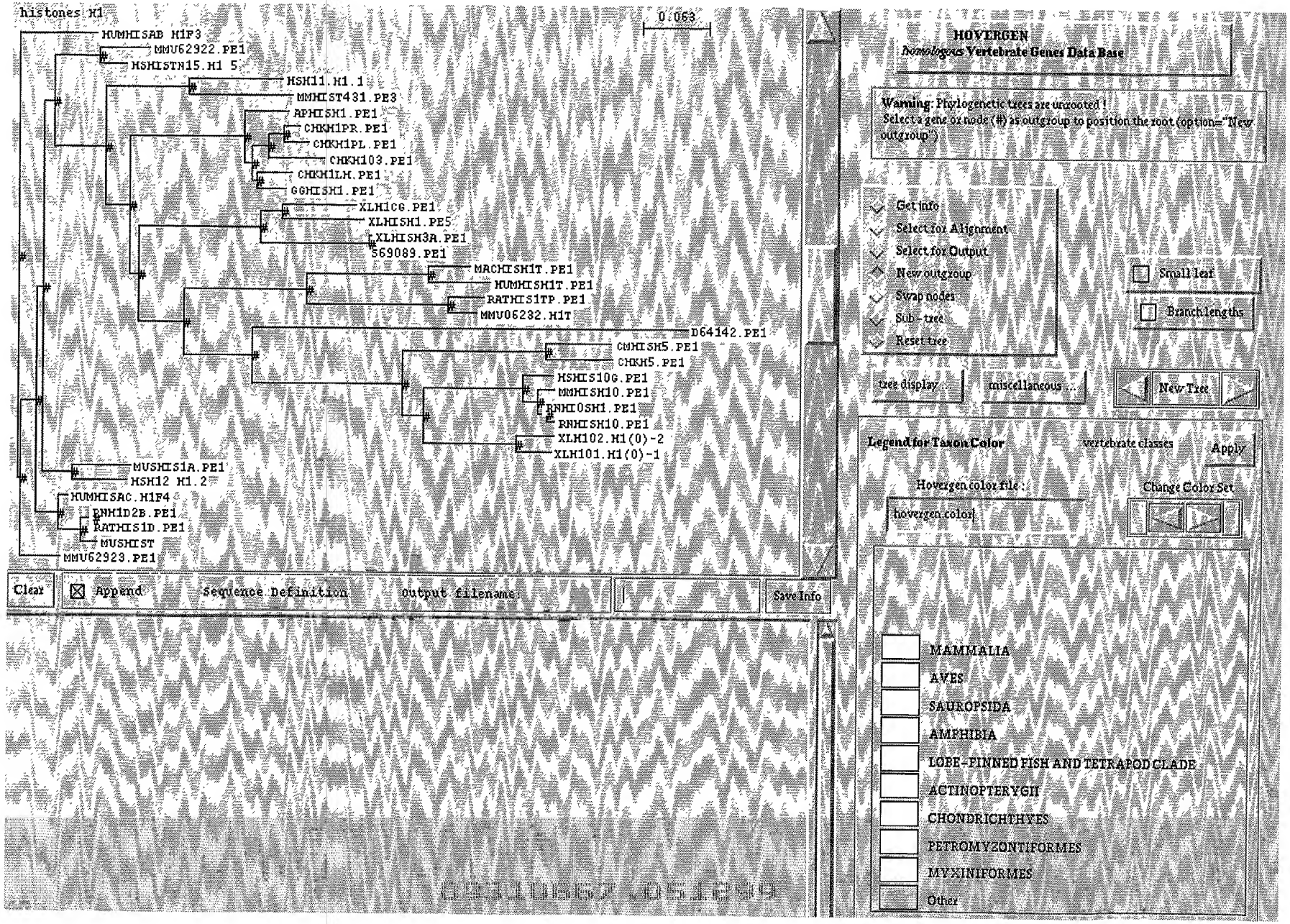

histones H1
0.063
HUMHISAB H1F3
MMU62922.PE1
HSHISTN15.H1 5
HSH11.H1.1
MMHIST431.PE3
APHISH1.PE1
CHKH1PR.PE1
CHKH1PL.PE1
CHKH103.PE1
CHKH1LH.PE1
GGHISH1.PE1
XLH1C6.PE1
XLHISH1.PE5
XLHISH3A.PE1
569089.PE1
MACHISH1T.PE1
HUMHISH1T.PE1
RATHIS1TP.PE1
MMV06232.H1T
D64142.PE1
CMHISH5.PE1
CHKH5.PE1
HSHIS10G.PE1
MMHISH10.PE1
RNHI0SH1.PE1
RNHISH10.PE1
XLH102.H1(0)-2
XLH101.H1(0)-1
MUSHIS1A.PE1
HSH12 H1.2
HUMHISAC.H1F4
RNH1D2B.PE1
RATHIS1D.PE1
MUSHIST
MMV62923.PE1
Clear
Append
Sequence Definition
Output filename:
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted !
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display...
miscellaneous...
New Tree
Legend for Taxon Color
vertebrate classes
Apply
Hovergen color file:
hovergen.color
Change Color Set
MAMMALIA
AVES
SAUROPSIDA
AMPHIBIA
LOBE-FINNED FISH AND TETRAPOD CLADE
ACTINOPTERYGII
CHONDRICHTHYES
PETROMYZONTIFORMES
MYXINIFORMES
Other

Figure 31

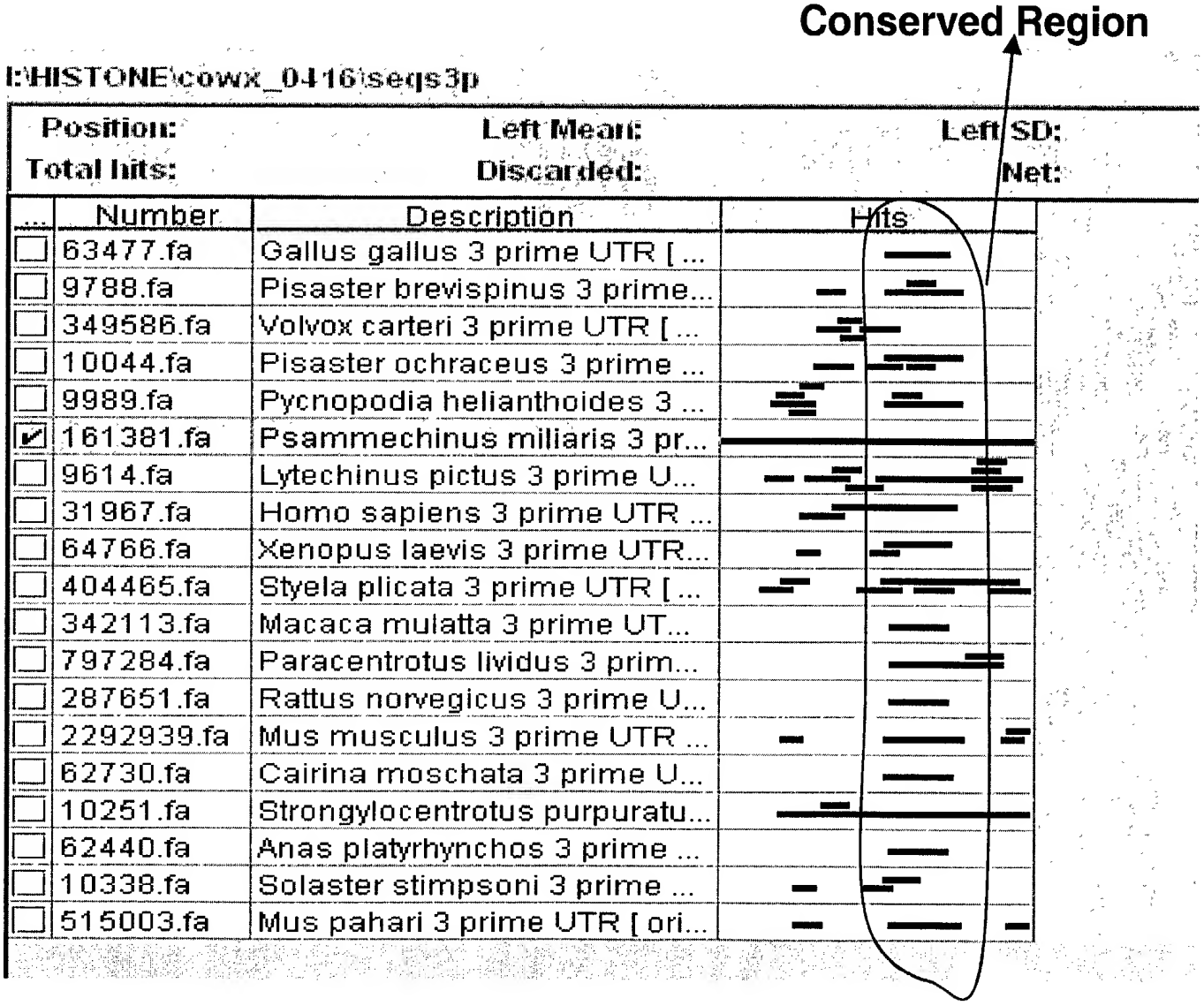
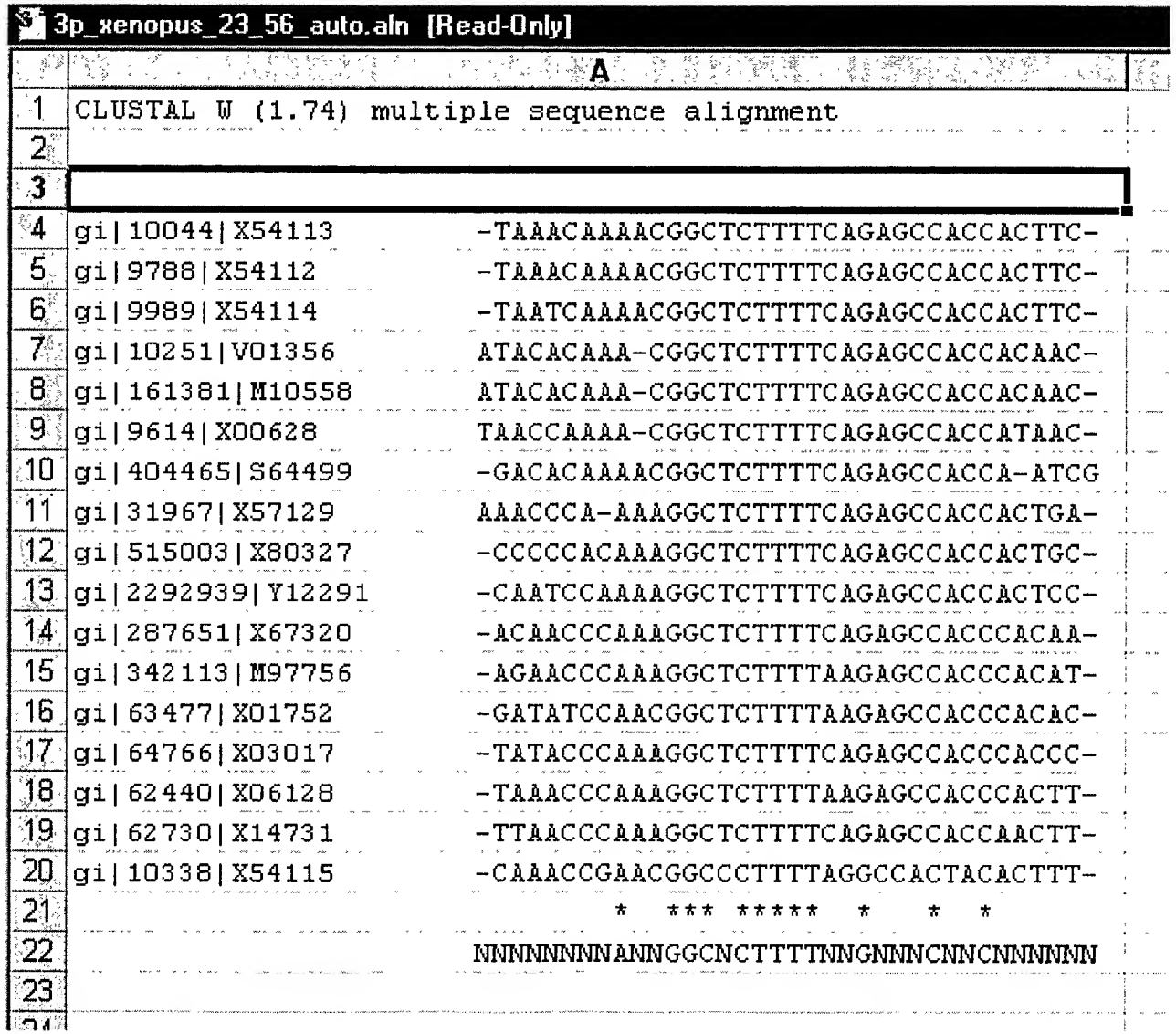


Figure 32



3p_xenopus_23_56_auto.aln [Read-Only]

	A
1	CLUSTAL W (1.74) multiple sequence alignment
2	
3	
4	gi|10044|X54113 -TAAACAAAACGGCTCTTTTCAGAGCCACCACTTC-
5	gi|9788|X54112 -TAAACAAAACGGCTCTTTTCAGAGCCACCACTTC-
6	gi|9989|X54114 -TAATCAAAACGGCTCTTTTCAGAGCCACCACTTC-
7	gi|10251|V01356 ATACACAAA-CGGCTCTTTTCAGAGCCACCACAAC-
8	gi|161381|M10558 ATACACAAA-CGGCTCTTTTCAGAGCCACCACAAC-
9	gi|9614|X00628 TAACCAAAA-CGGCTCTTTTCAGAGCCACCATAAC-
10	gi|404465|S64499 -GACACAAAACGGCTCTTTTCAGAGCCACCA-ATCG
11	gi|31967|X57129 AAACCCA-AAAGGCTCTTTTCAGAGCCACCACTGA-
12	gi|515003|X80327 -CCCCCACAAAGGCTCTTTTCAGAGCCACCACTGC-
13	gi|2292939|Y12291 -CAATCCAAAAGGCTCTTTTCAGAGCCACCACTCC-
14	gi|287651|X67320 -ACAACCCAAAGGCTCTTTTCAGAGCCACCCACAA-
15	gi|342113|M97756 -AGAACCCAAAGGCTCTTTTAAGAGCCACCCACAT-
16	gi|63477|X01752 -GATATCCAACGGCTCTTTTAAGAGCCACCCACAC-
17	gi|64766|X03017 -TATACCCAAAGGCTCTTTTCAGAGCCACCCACCC-
18	gi|62440|X06128 -TAAACCCAAAGGCTCTTTTAAGAGCCACCCACTT-
19	gi|62730|X14731 -TTAACCCAAAGGCTCTTTTCAGAGCCACCAACTT-
20	gi|10338|X54115 -CAAACCGAACGGCCCTTTTAGGCCACTACACTTT-
21	* *** ***** * * *
22	NNNNNNNNANNGGCNCTTTTNNGNNNCNNCNNNNNN
23	

Figure 33

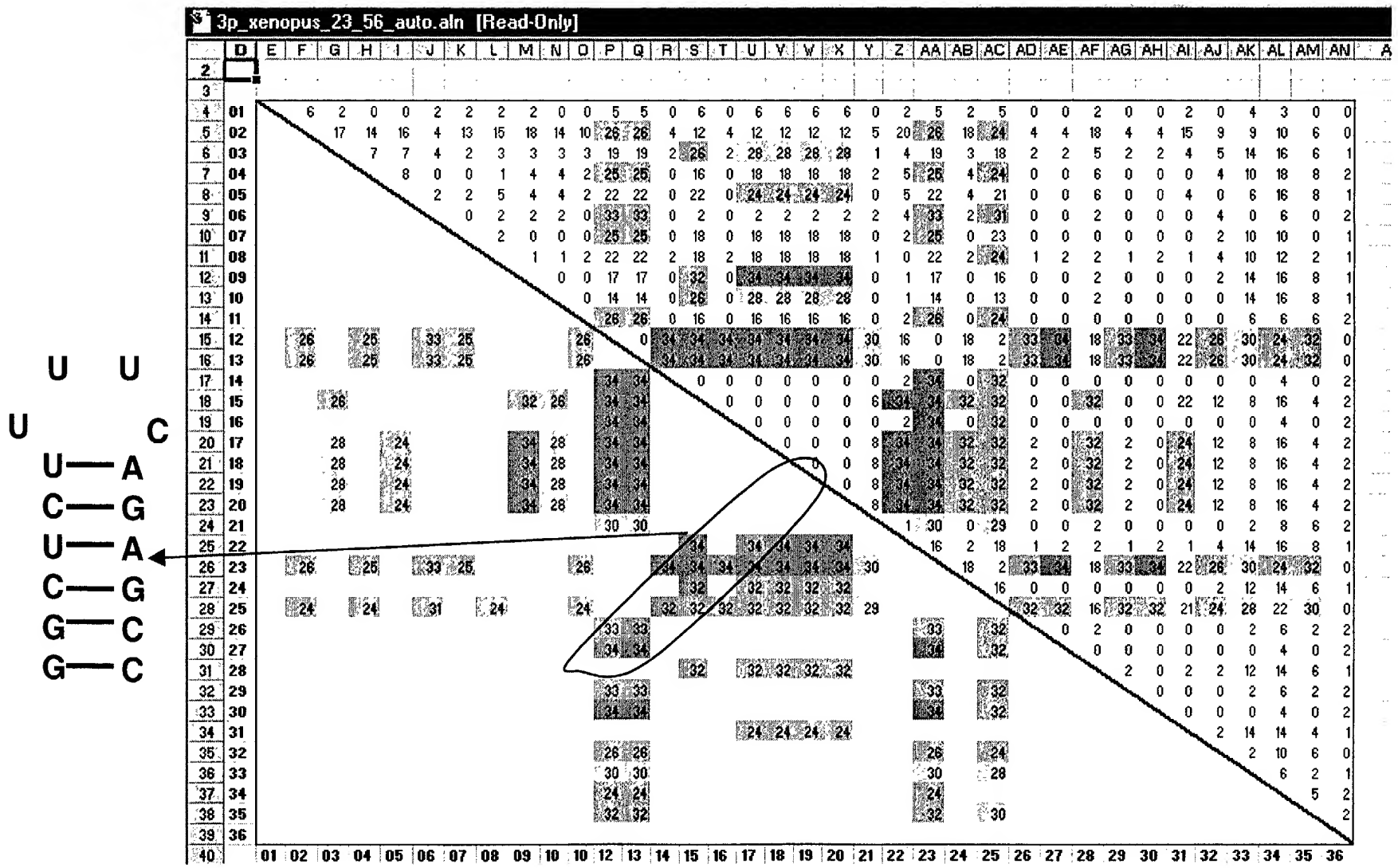


Figure 34

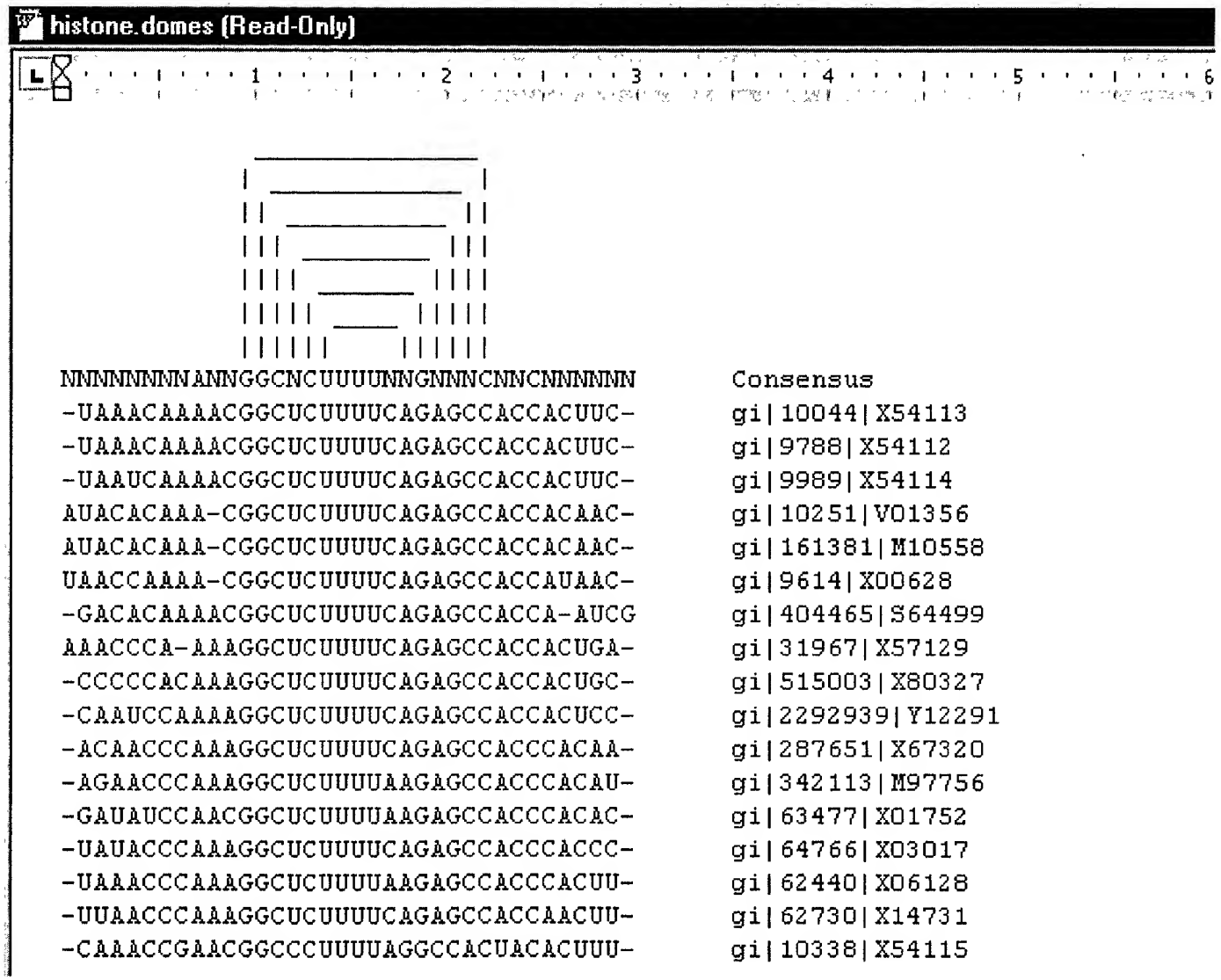


Figure 35

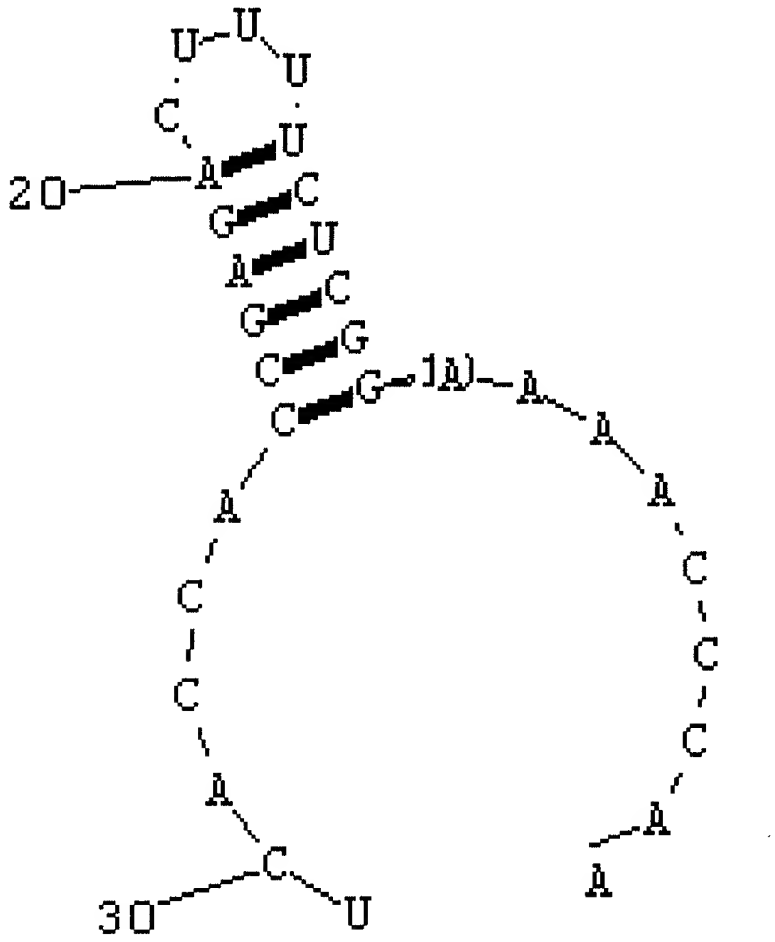


Figure 36

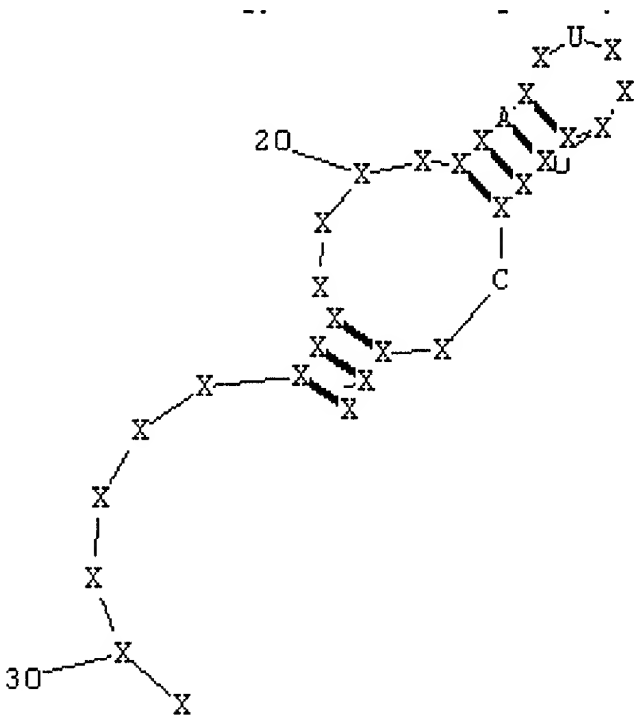


Figure 37

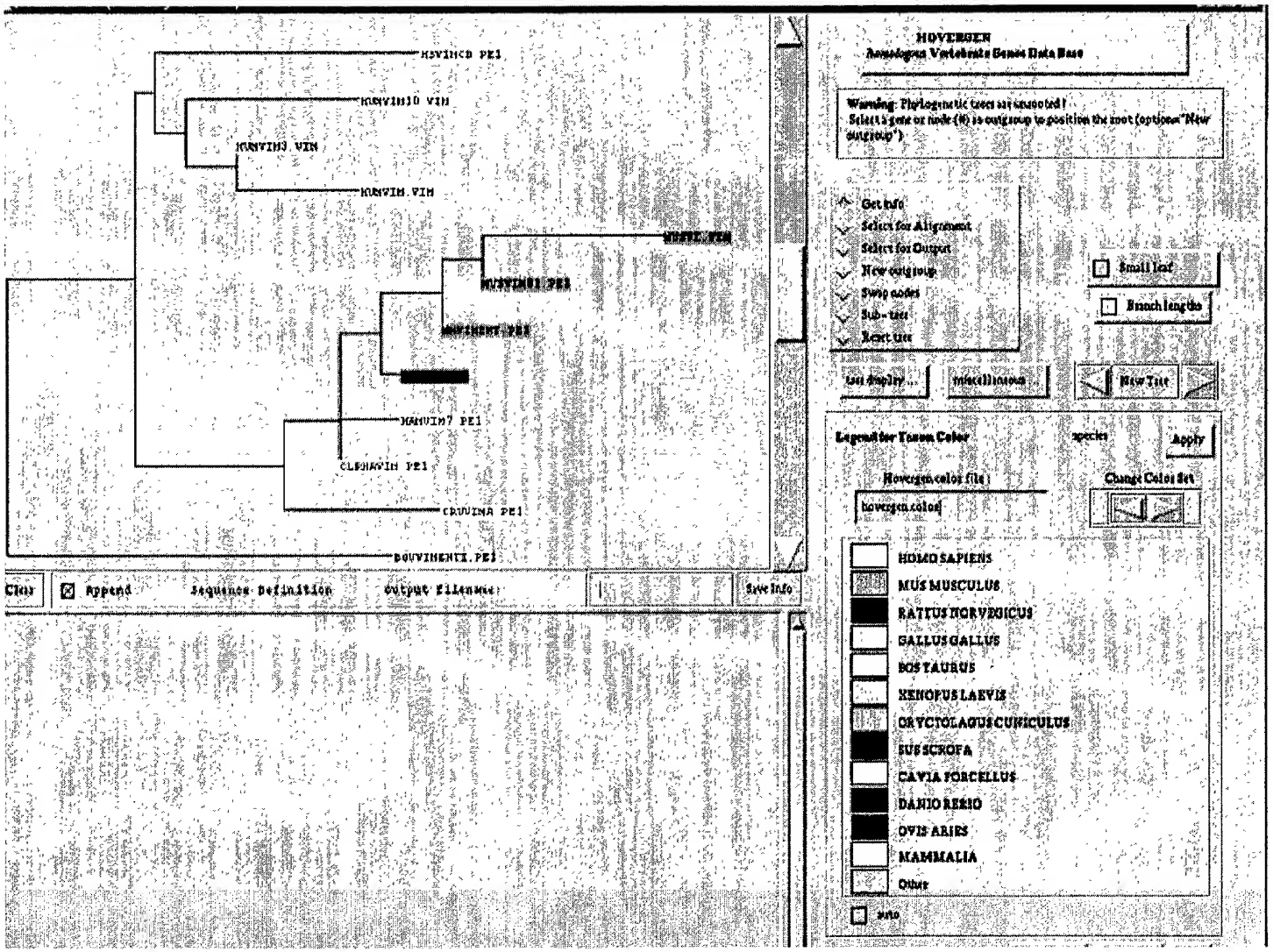


Figure 38

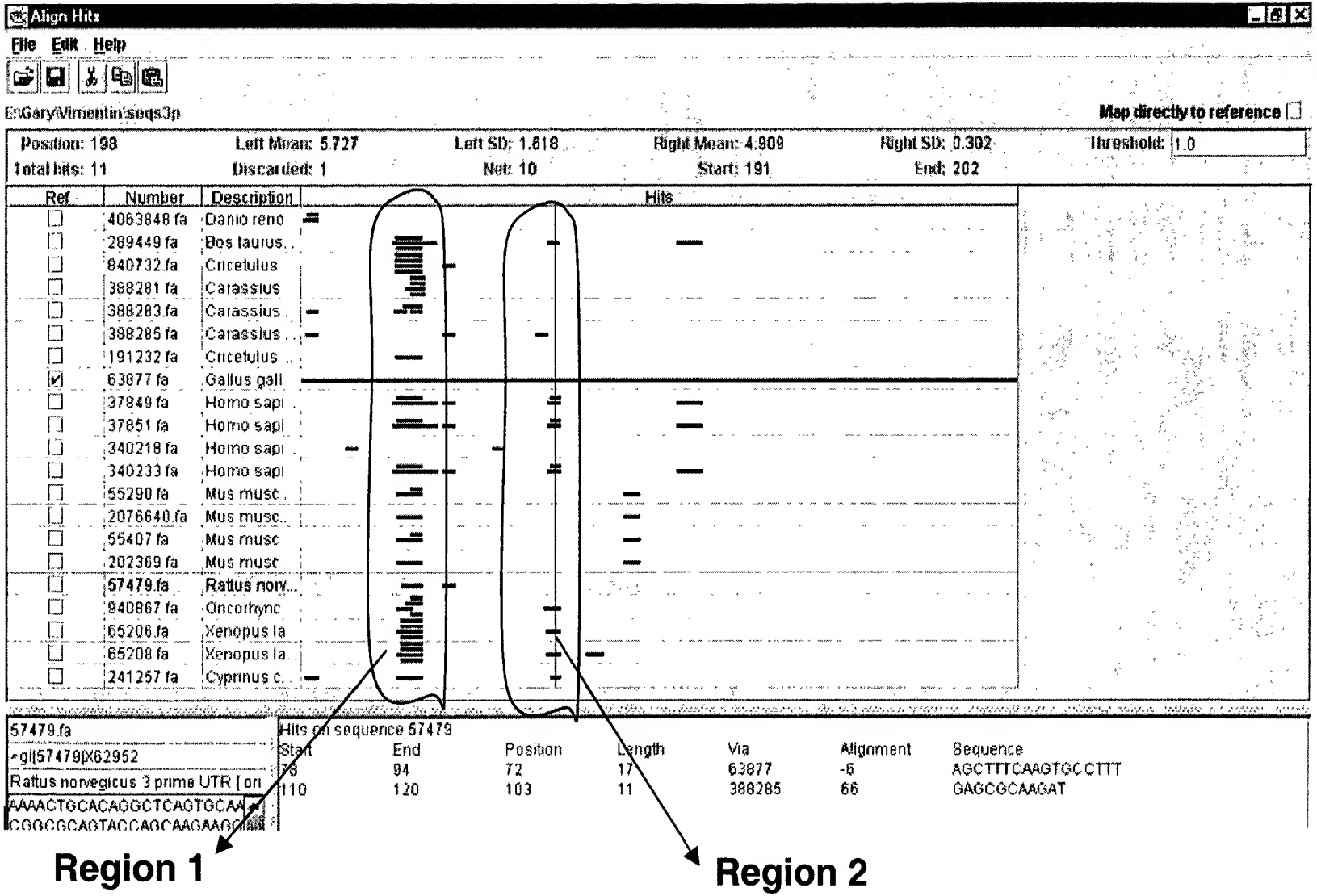


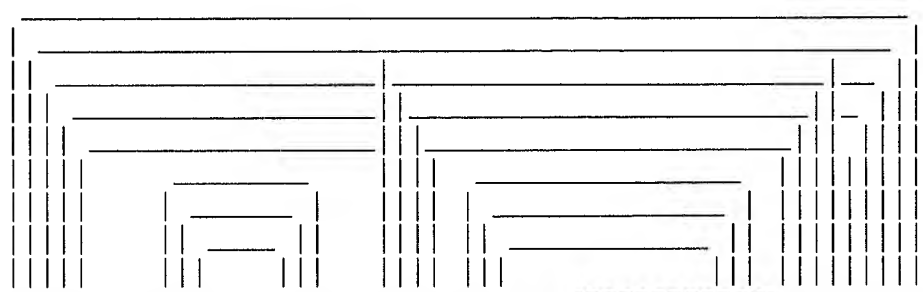
Figure 39

```
CLUSTAL W (1.74) multiple sequence alignment


gi|191232|M16718        TATCTTAAGGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGTAA
gi|202369|M26251        TATCTTAGGAAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|2076640|Y07738       TATCTTAGGAAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|289449|L13263        TATCTTAAAGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCG-AAGATA
gi|340233|M25246        TATCTTAAAGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|37849|X56134         TATCTTAAAGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|37851|Z19554         TATCTTAAAGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|388281|L23840        CAACCCACAATAACTGCTTCAAAGTGCCTTTCTGCACAGAAATA---GCGTTTGAGC
gi|388283|L23842        CTACCCACAATAACTGCTTCAAAGTGCCTTTCTGC-CAGAAGTACAAGCATTTGAGC
gi|55290|X51438         TATCTTAGGAAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|55407|X56397         TATCTTAGGAAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
gi|57479|X62952         TATCTTAGAAAAAAAGCTTTCAAGTGCCTTTCTGCAGTTTT-CAGGAGCGCAAGATA
gi|63877|V00447         TGTCTTAAAGGAAGAGCTTTCAAGTGCCTTTCTCCAGTTTTTCATGAGCGCAAGATT
gi|65206|X16843         ACTTTGAAGAAAACAGCTTTCAAGTGCCTTT-TGCAGTCAATGGAGAGCGCAAGATA
gi|65208|X16844         AATTTGAAGAAAACAGCTTTCAAGTGCCTTT-TGCAGTTAATGGAGAGCGCAAGATA
gi|840732|X87227        TATCTTAAGGAAACAGCTTTCAAGTGCCTTTCTGCAGTTTTTCAGGAGCGCAAGATA
```

Figure 40

Score: 445.0



```
NNNNNNANNNNAANNGCUUNNAAGUGCCUUUNUNCNNNNNNNNNNNNNGCNNNNGNNN  Consensus
UAUCUUAAGGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGUAA  gi|191232|M16718
UAUCUUAGGAAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|202369|M26251
UAUCUUAGGAAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|2076640|Y07738
UAUCUUAAAGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCG-AAGAUA  gi|289449|L13263
UAUCUUAAAGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|340233|M25246
UAUCUUAAAGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|37849|X56134
UAUCUUAAAGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|37851|Z19554
CAACCCACAAUAACUGCUUCAAAGUGCCUUUCUGCACAGAAAUA---GCGUUUGAGC  gi|388281|L23840
CUACCCACAAUAACUGCUUCAAAGUGCCUUUCUGC-CAGAAGUACAAGCAUUUGAGC  gi|388283|L23842
UAUCUUAGGAAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|55290|X51438
UAUCUUAGGAAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|55407|X56397
UAUCUUAGAAAAAAAGCUUUCAAGUGCCUUUCUGCAGUUUU-CAGGAGCGCAAGAUA  gi|57479|X62952
UGUCUUAAAGGAAGAGCUUUCAAGUGCCUUUCUCCAGUUUUUCAUGAGCGCAAGAUU  gi|63877|V00447
ACUUUGAAGAAAACAGCUUUCAAGUGCCUUU-UGCAGUCAAUGGAGAGCGCAAGAUA  gi|65206|X16843
AAUUUGAAGAAAACAGCUUUCAAGUGCCUUU-UGCAGUUAAUGGAGAGCGCAAGAUA  gi|65208|X16844
UAUCUUAAGGAAACAGCUUUCAAGUGCCUUUCUGCAGUUUUUCAGGAGCGCAAGAUA  gi|840732|X87227
```

Figure 41

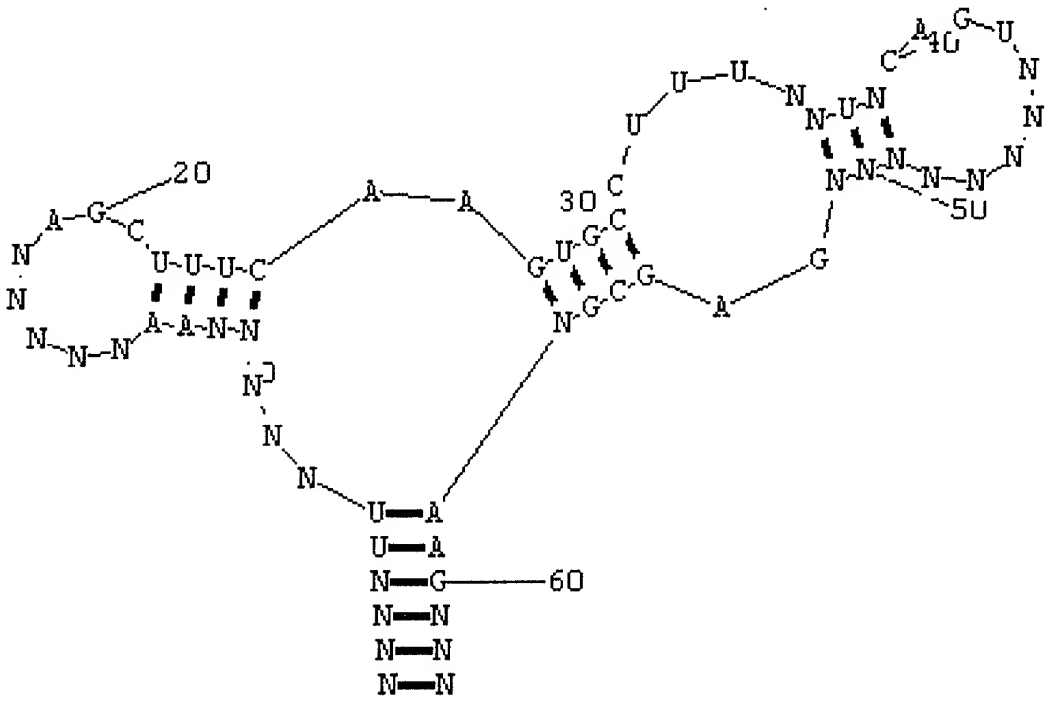


Figure 42

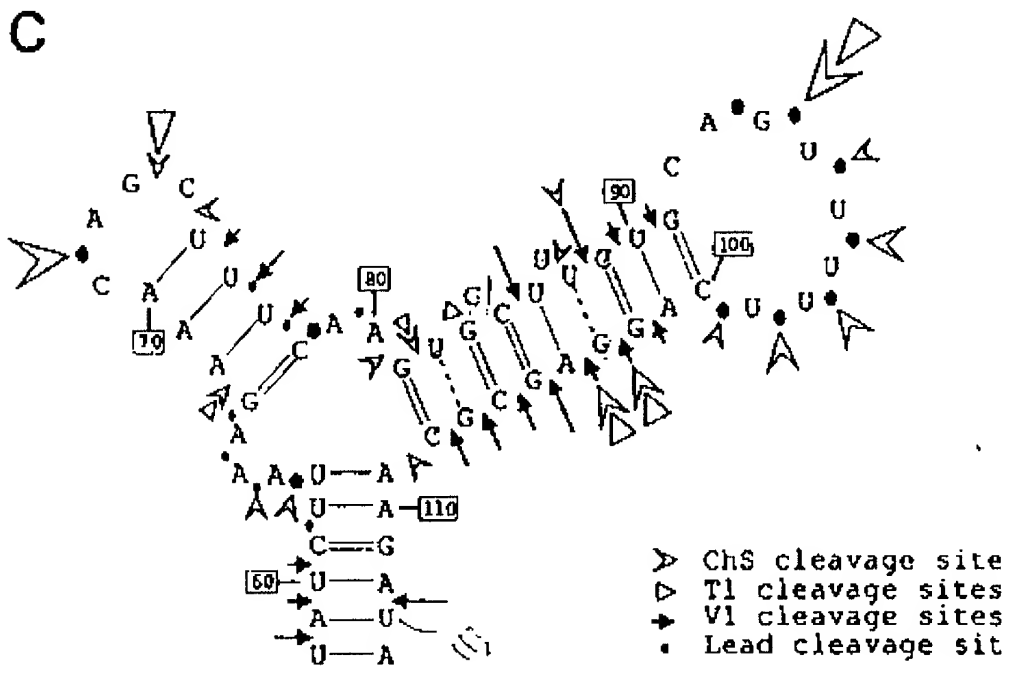
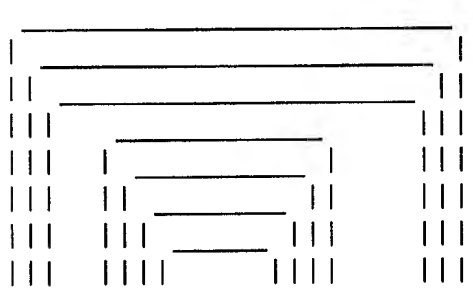


Figure 43

```
CLUSTAL W (1.74) multiple sequence alignment


gi|241257|S76850        ACCACGATGT-CTGTAGTTTACACTGTTGAA
gi|289449|L13263        TTTACAACATAATCTAGTTTACCGAAGACGC
gi|340233|M25246        TTTACAACATAATCTAGTTTACAGAAAAATC
gi|37849|X56134         TTTACAACATAATCTAGTTTACAGAAAAATC
gi|37851|Z19554         TTTACAACATAATCTAGTTTACAGAAAAATC
gi|388285|L23841        TCCACACTGGAGTAAAACGAGGAAGAATGAA
gi|63877|V00447         TTTACAATGGAGTCTAGTTTACAAATAGCAA
gi|65206|X16843         GCTTCCTTCT-GTCTAGTTTACAGACTGTAA
gi|65208|X16844         GCTTCCTTCT-GTCTAGTTTACAGACTATGT
gi|940867|Z50738        ACCACACTGA-GTCTAGTTTACACTTGCGCT
```

Figure 44



```
NNNNCNNNNNNNNUNNANNNNNNNNNNNNNNNN      Consensus
ACCACGAUGU-CUGUAGUUUACACUGUUGAA        gi|241257|S76850
UUUACAACAUAAUCUAGUUUACCGAAGACGC        gi|289449|L13263
UUUACAACAUAAUCUAGUUUACAGAAAAAUC        gi|340233|M25246
UUUACAACAUAAUCUAGUUUACAGAAAAAUC        gi|37849|X56134
UUUACAACAUAAUCUAGUUUACAGAAAAAUC        gi|37851|Z19554
UCCACACUGGAGUAAAAACGAGGAAGAAUGAA       gi|388285|L23841
UUUACAAUGGAGUCUAGUUUACAAAUAGCAA        gi|63877|V00447
GCUUCCUUCU-GUCUAGUUUACAGACUGUAA        gi|65206|X16843
GCUUCCUUCU-GUCUAGUUUACAGACUAUGU        gi|65208|X16844
ACCACACUGA-GUCUAGUUUACACUUGCGCU        gi|940867|Z50738
```

Figure 45

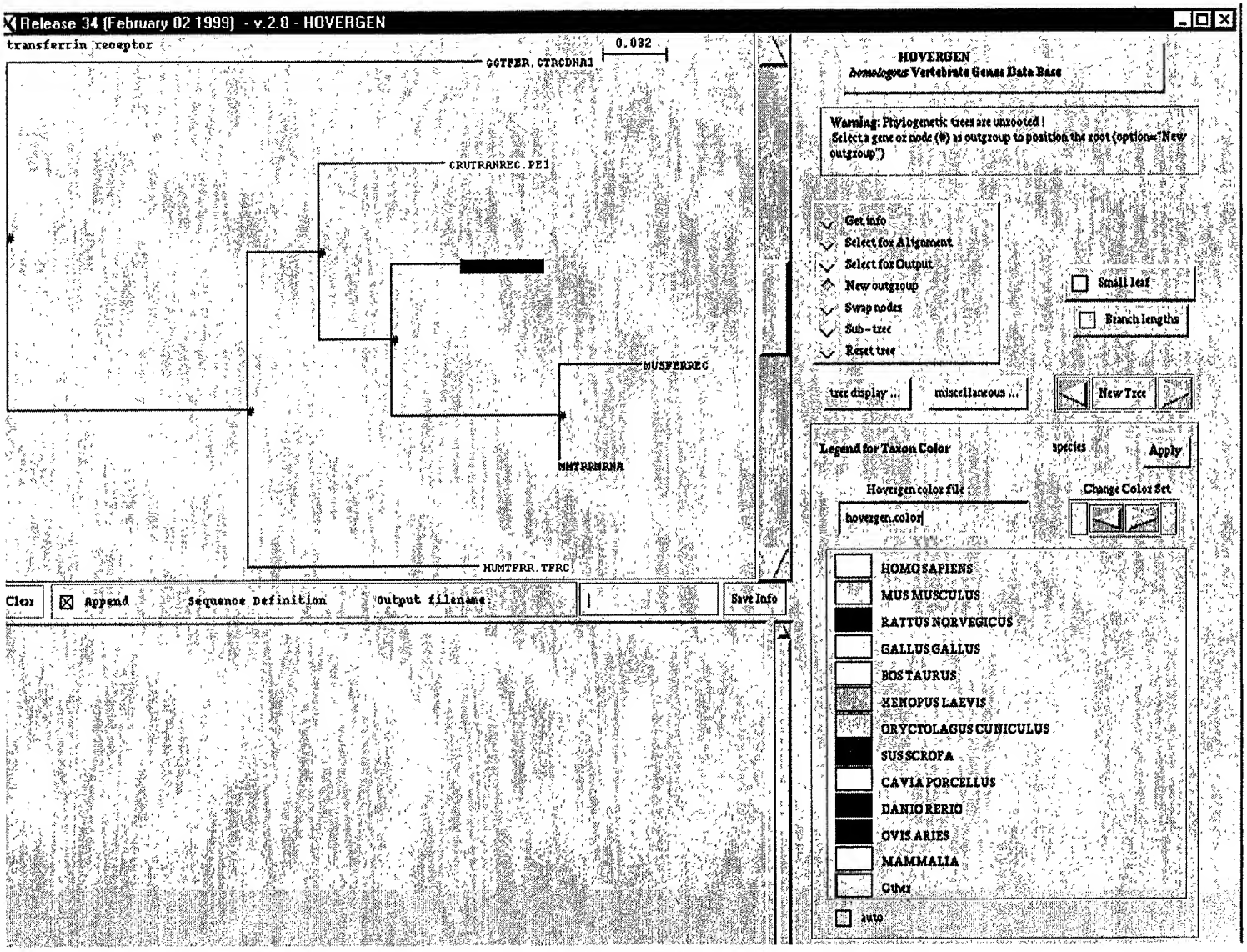


Figure 46

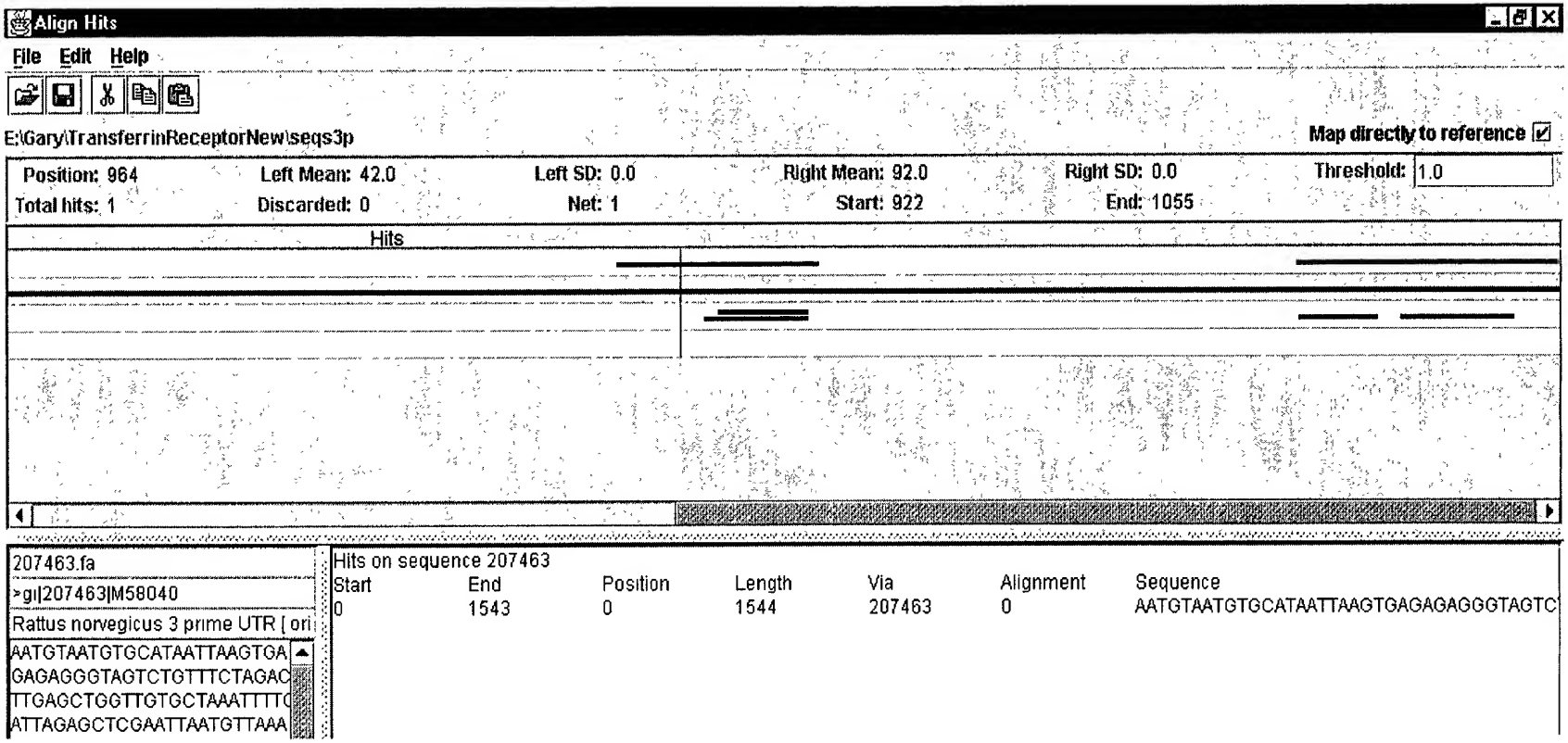

Align Hits
File Edit Help
E:\Gary\TransferrinReceptorNew\seqs3p
Map directly to reference
Position: 964
Left Mean: 42.0
Left SD: 0.0
Right Mean: 92.0
Right SD: 0.0
Threshold: 1.0
Total hits: 1
Discarded: 0
Net: 1
Start: 922
End: 1055
Hits
207463.fa
>gi|207463|M58040
Rattus norvegicus 3 prime UTR [ori
AATGTAATGTGCATAATTAAGTGA
GAGAGGGTAGTCTGTTTCTAGAC
TTGAGCTGGTTGTGCTAAATTTTC
ATTAGAGCTCGAATTAATGTTAAA
Hits on sequence 207463
Start
End
Position
Length
Via
Alignment
Sequence
0
1543
0
1544
207463
0
AATGTAATGTGCATAATTAAGTGAGAGAGGGTAGTC

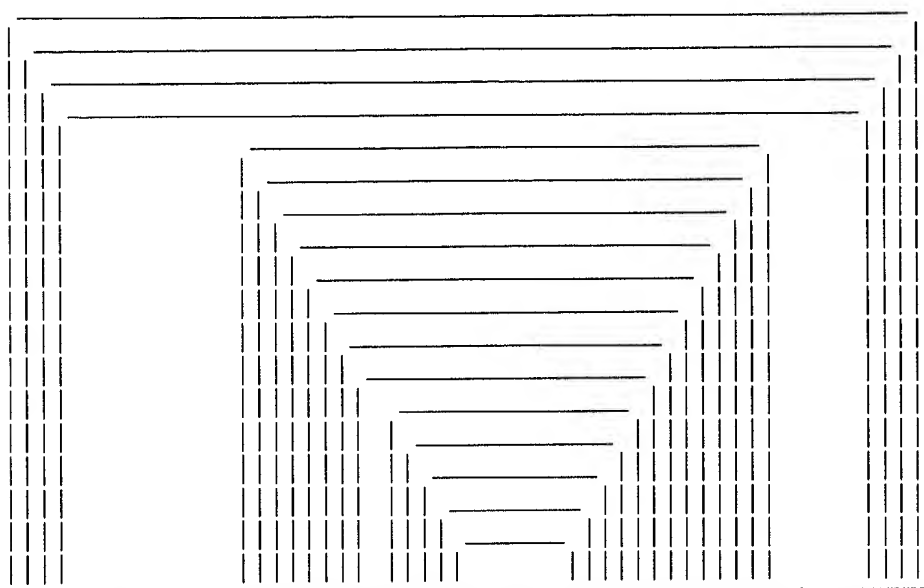
Figure 47

```
CLUSTAL W (1.74) multiple sequence alignment


gi|207463|M58040
GTTTTTGGCACTGAGATATTTATTGTTTATTTATCAGTGACAGAGTTCACTATAAATAGTGTTTTTTTAAT
gi|37432|X01060
CCTTTTGGCACTGAGATATTTATTGTTTATTTATCAGTGACAGAGTTCACTATAAATGGTGTTTTTTTAAT
gi|63357|X13753        ----------
TGAGATATTTATTTTTTATTTATCAGTGACAGCGTTCACTATAAATGGTGTTTTTTTTAT
```

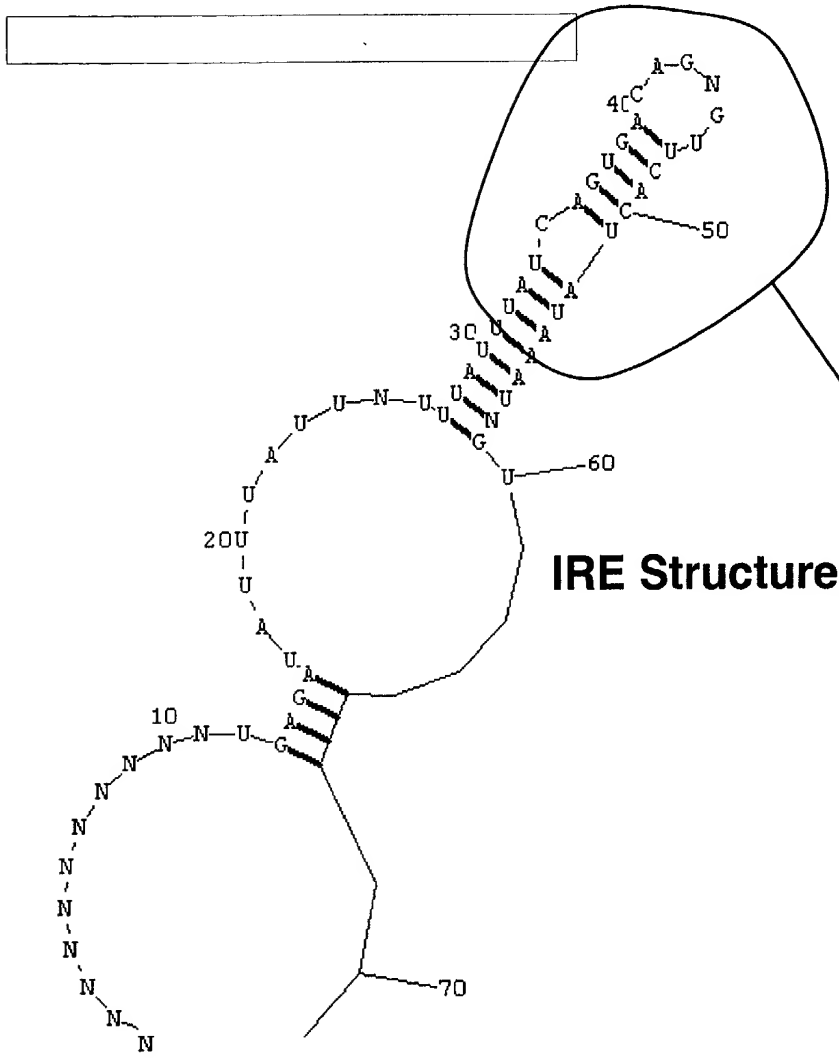
Figure 48

Score: 102.0



```
NNNNNNNNNNNNUGAGAUAUUUAUUNUUUAUUUAUCAGUGACAGNGUUCACUAUAAAUNGUGUUUUUUUUNAU Consensus
GUUUUUGGCACUGAGAUAUUUAUUGUUUAUUUAUCAGUGACAGAGUUCACUAUAAAUAGUGUUUUUUUAAU gi|207463|
CCUUUUGGCACUGAGAUAUUUAUUGUUUAUUUAUCAGUGACAGAGUUCACUAUAAAUGGUGUUUUUUUAAU gi|37432|X
-----------UGAGAUAUUUAUUUUUUAUUUAUCAGUGACAGCGUUCACUAUAAAUGGUGUUUUUUUUAU gi|63357|X
```

Figure 49



IRE Structure predicted by NMR

Figure 50

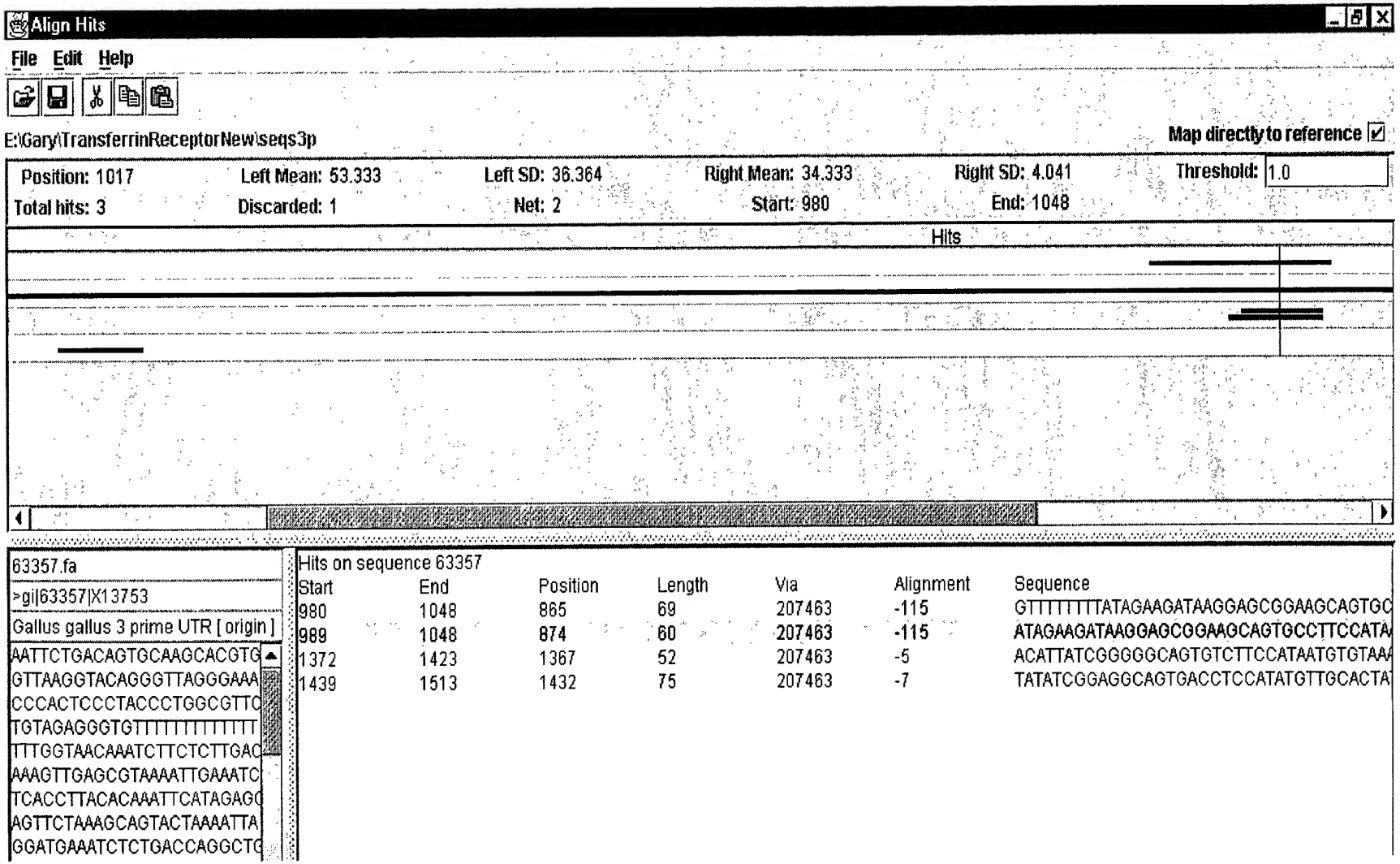


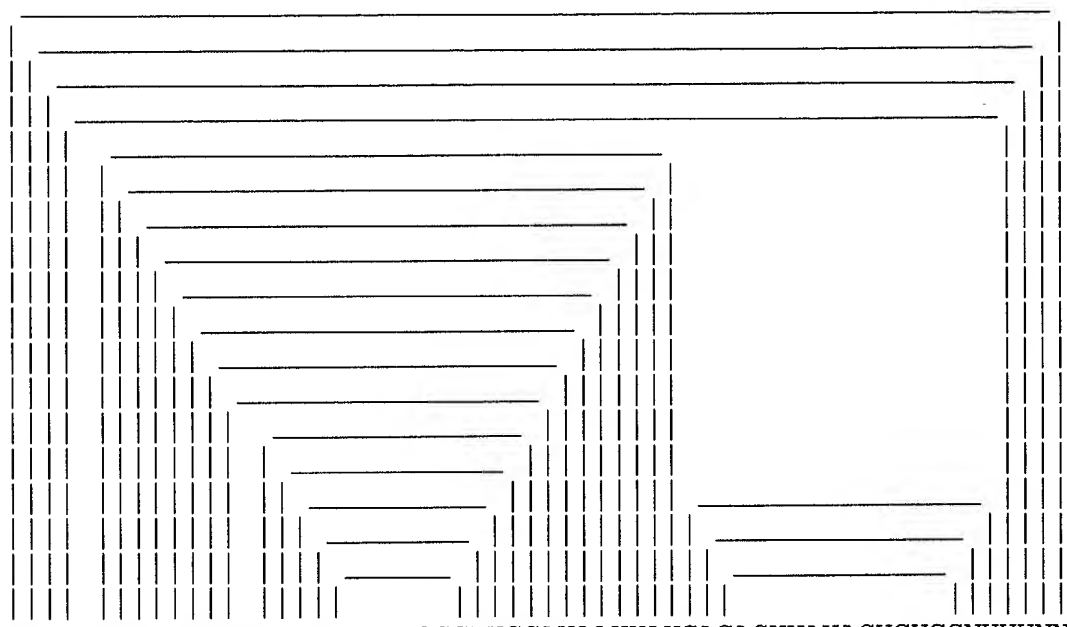
Figure 51

```
CLUSTAL W (1.74) multiple sequence alignment


gi|207463|M58040
TAGAAGATAATTATCGGAAGCAGTGCCTTCCATAATTATGACAGTTATACTGTCGTTTTCT
gi|37432|X01060
TAGAATATAATTATCGGAAGCAGTGCCTTCCATAATTATGACAGTTATACTGTCGGTTTTT
gi|63357|X13753
TAGAAGATAAGGAGCGGAAGCAGTGCCTTCCATAATTATGACAGTTATACTGTCGTTTTTG
```

Figure 52

Score: 115.0



UAGAANAUAANNANCGGAAGCAGUGCCUUCCAUAAUUAUGACAGUUAUACUGUCGNUUUNN	Consensus
UAGAAGAUAAUUAUCGGAAGCAGUGCCUUCCAUAAUUAUGACAGUUAUACUGUCGUUUUCU	gi|207463|M58040
UAGAAUAUAAUUAUCGGAAGCAGUGCCUUCCAUAAUUAUGACAGUUAUACUGUCGGUUUUU	gi|37432|X01060
UAGAAGAUAAGGAGCGGAAGCAGUGCCUUCCAUAAUUAUGACAGUUAUACUGUCGUUUUUG	gi|63357|X13753

Figure 53

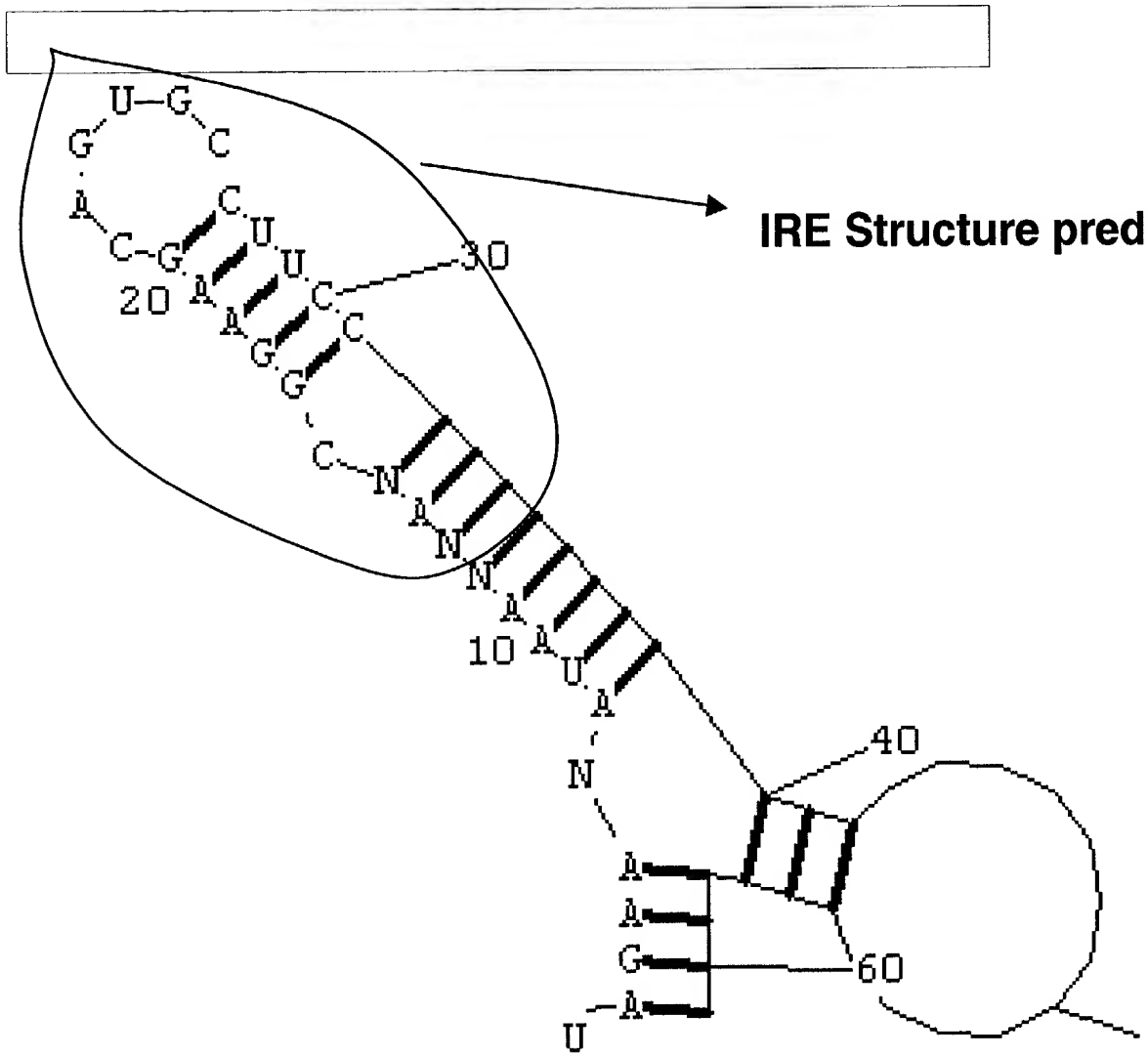


Figure 54

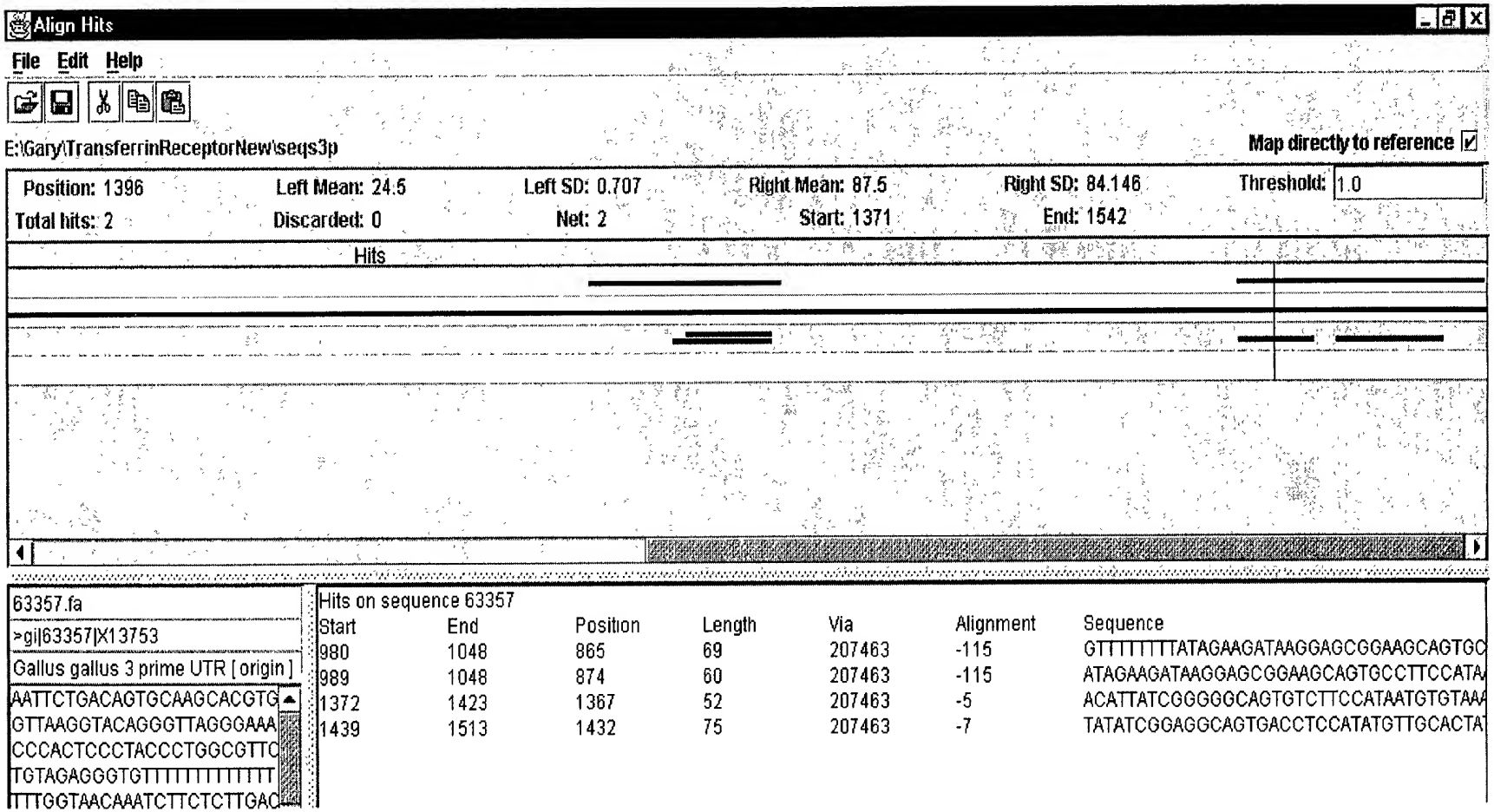


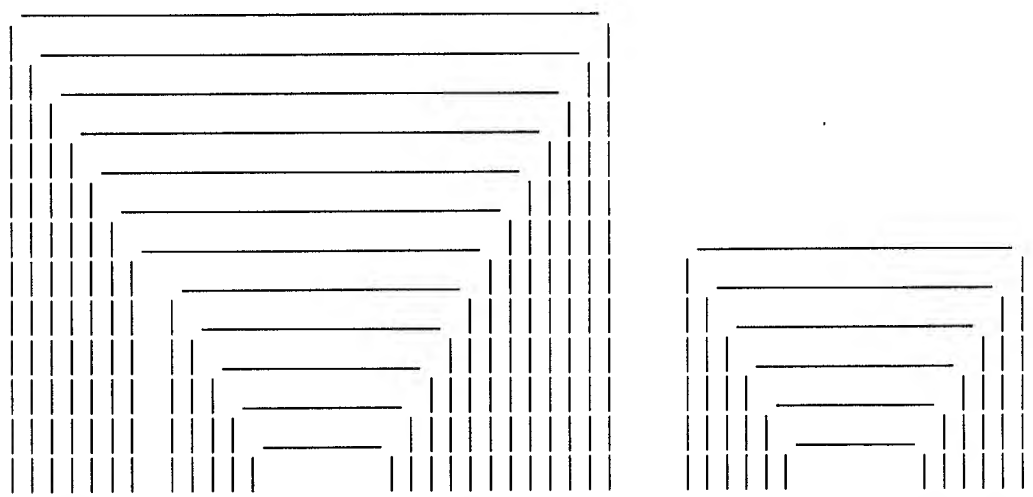
Figure 55

```
CLUSTAL W (1.74) multiple sequence alignment


gi|207463|M58040        ACATTATCGGGAGCAGTGTCTTCCATAATGTATAAAGAACAAGGTAGTTTTT
gi|37432|X01060         ACATTATCGGGAGCAGTGTCTTCCATAATGTATAAAGAACAAGGTAGTTTTT
gi|63357|X13753          ACATTATCGGGGGCAGTGTCTTCCATAATGTGTAAAGAACAAGGTAGTTTTT
                         *********** ******************* *******************
```

Figure 56

Score: 108.0



```
ACAUUAUCGGGNGCAGUGUCUUCCAUAAUGUNUAAAGAACAAGGUAGUUUUU  Consensus
ACAUUAUCGGGAGCAGUGUCUUCCAUAAUGUAUAAAGAACAAGGUAGUUUUU  gi|207463|M58040
ACAUUAUCGGGAGCAGUGUCUUCCAUAAUGUAUAAAGAACAAGGUAGUUUUU  gi|37432|X01060
ACAUUAUCGGGGGCAGUGUCUUCCAUAAUGUGUAAAGAACAAGGUAGUUUUU  gi|63357|X13753
```

Figure 57

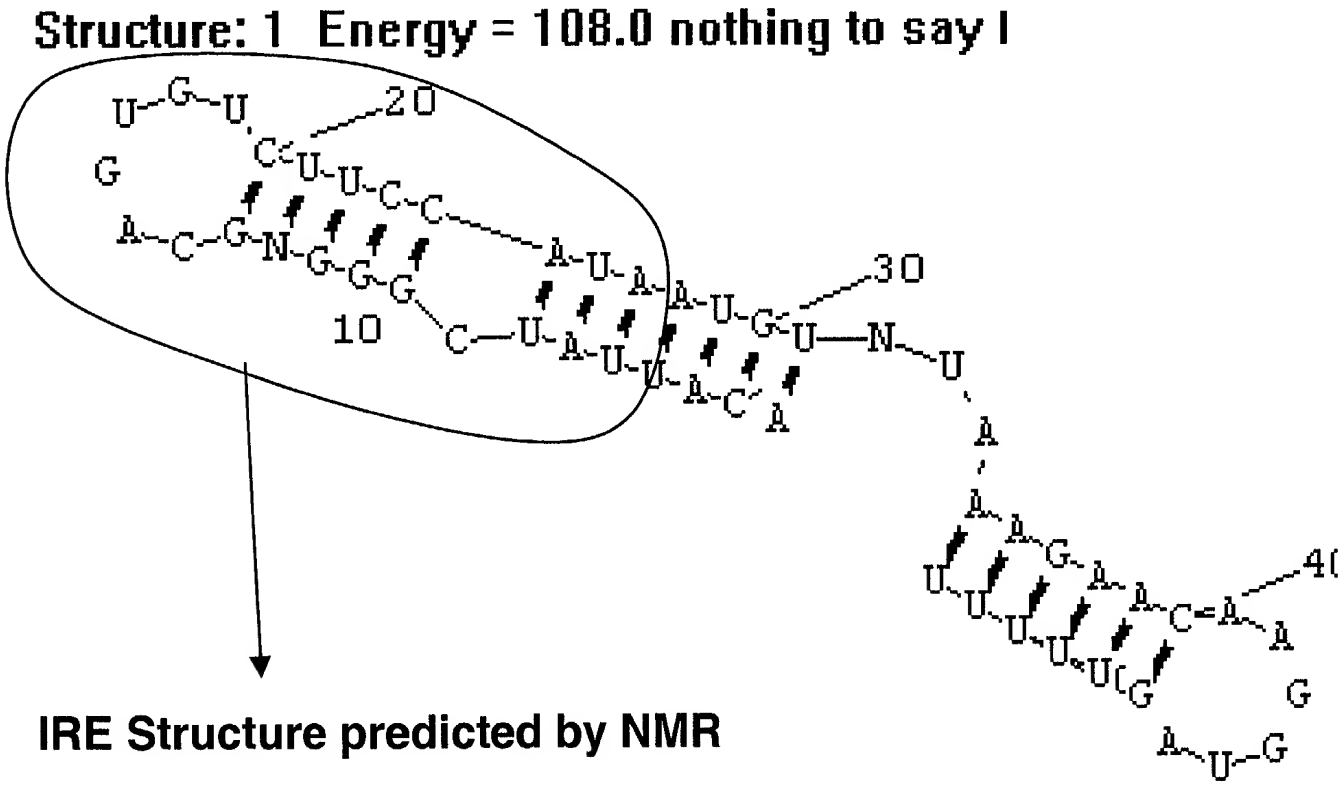


Figure 58

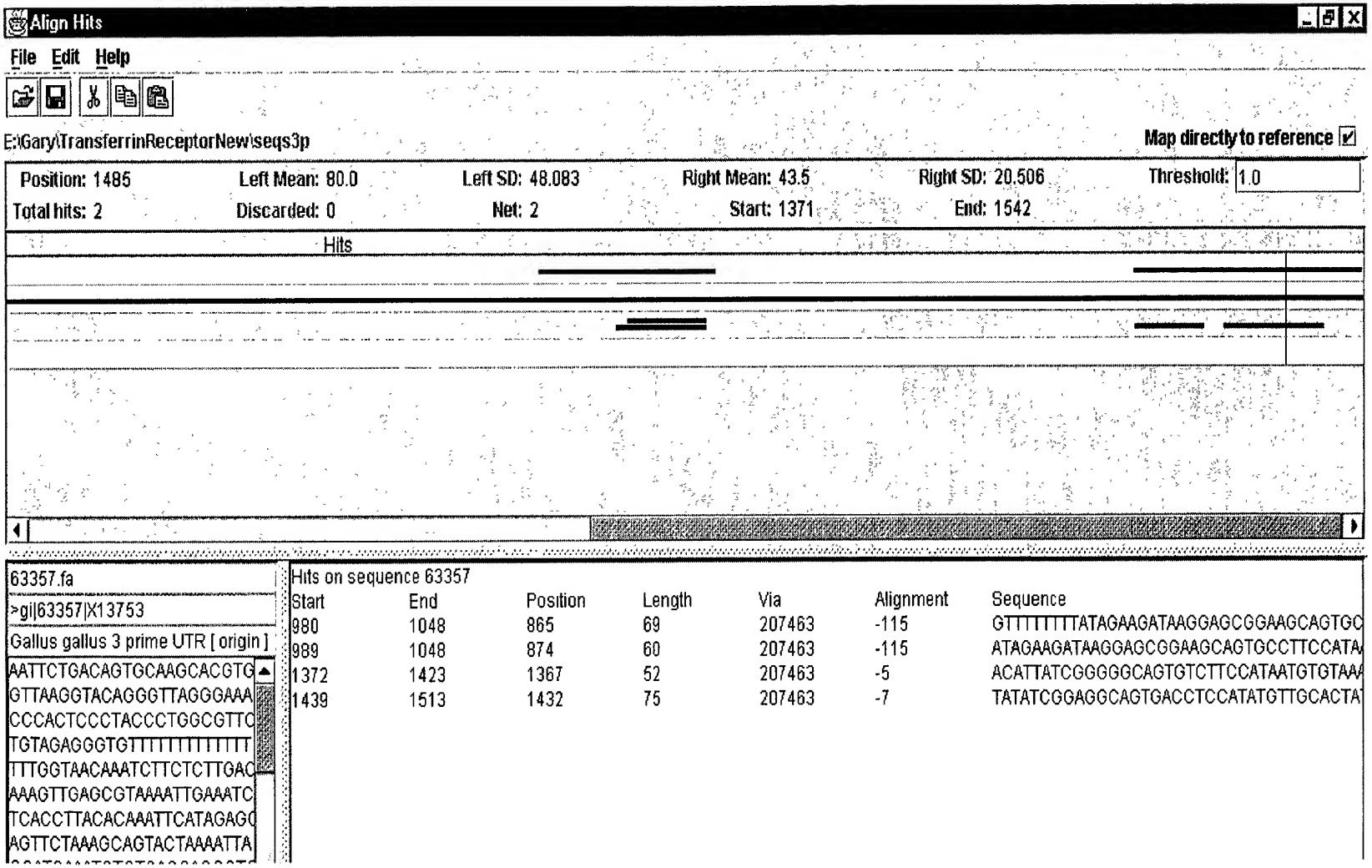


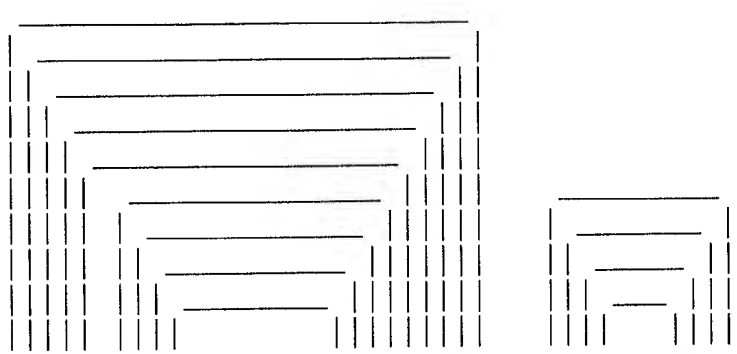
Figure 59

```
CLUSTAL W (1.74) multiple sequence alignment


gi|207463|M58040        TATATCGGAGACAGTGACCTCCATATGTTACACTAAGGGTG
gi|37432|X01060         TGTATCGGAGACAGTGATCTCCATATGTTACACTAAGGGTG
gi|63357|X13753          TATATCGGAGGCAGTGACCTCCATATGTTGCACTATGGGTG
                         * ******** ****** *********** ***** *****
```

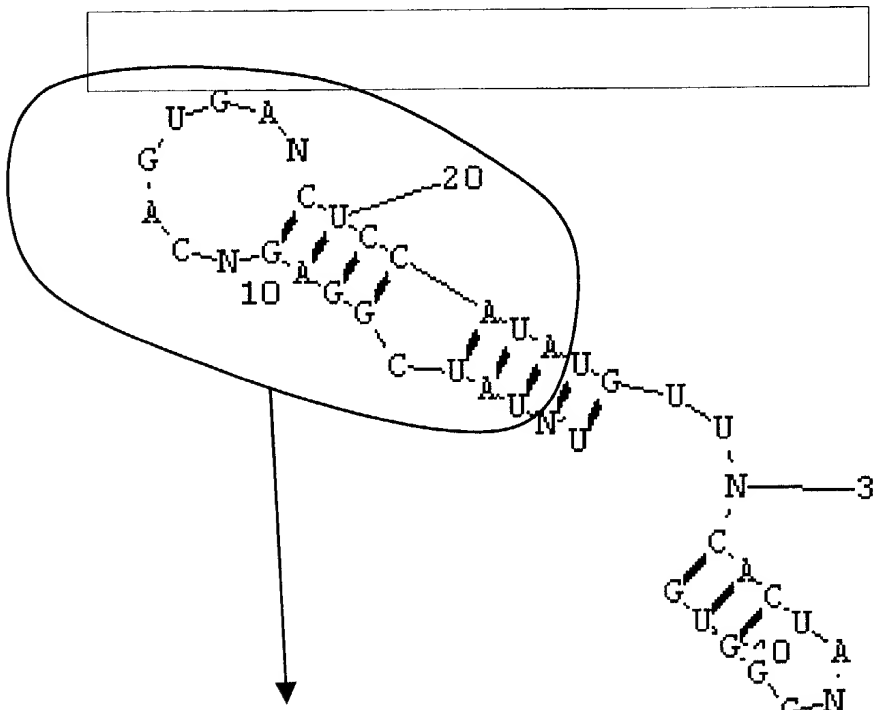
Figure 60

Score: 78.0



```
UNUAUCGGAGNCAGUGANCUCCAUAUGUUNCACUANGGGUG Consensus
UAUAUCGGAGACAGUGACCUCCAUAUGUUACACUAAGGGUG gi|207463|M58040
UGUAUCGGAGACAGUGAUCUCCAUAUGUUACACUAAGGGUG gi|37432|X01060
UAUAUCGGAGGCAGUGACCUCCAUAUGUUGCACUAUGGGUG gi|63357|X13753
```

Figure 61



IRE Structure predicted by NMR

Figure 62

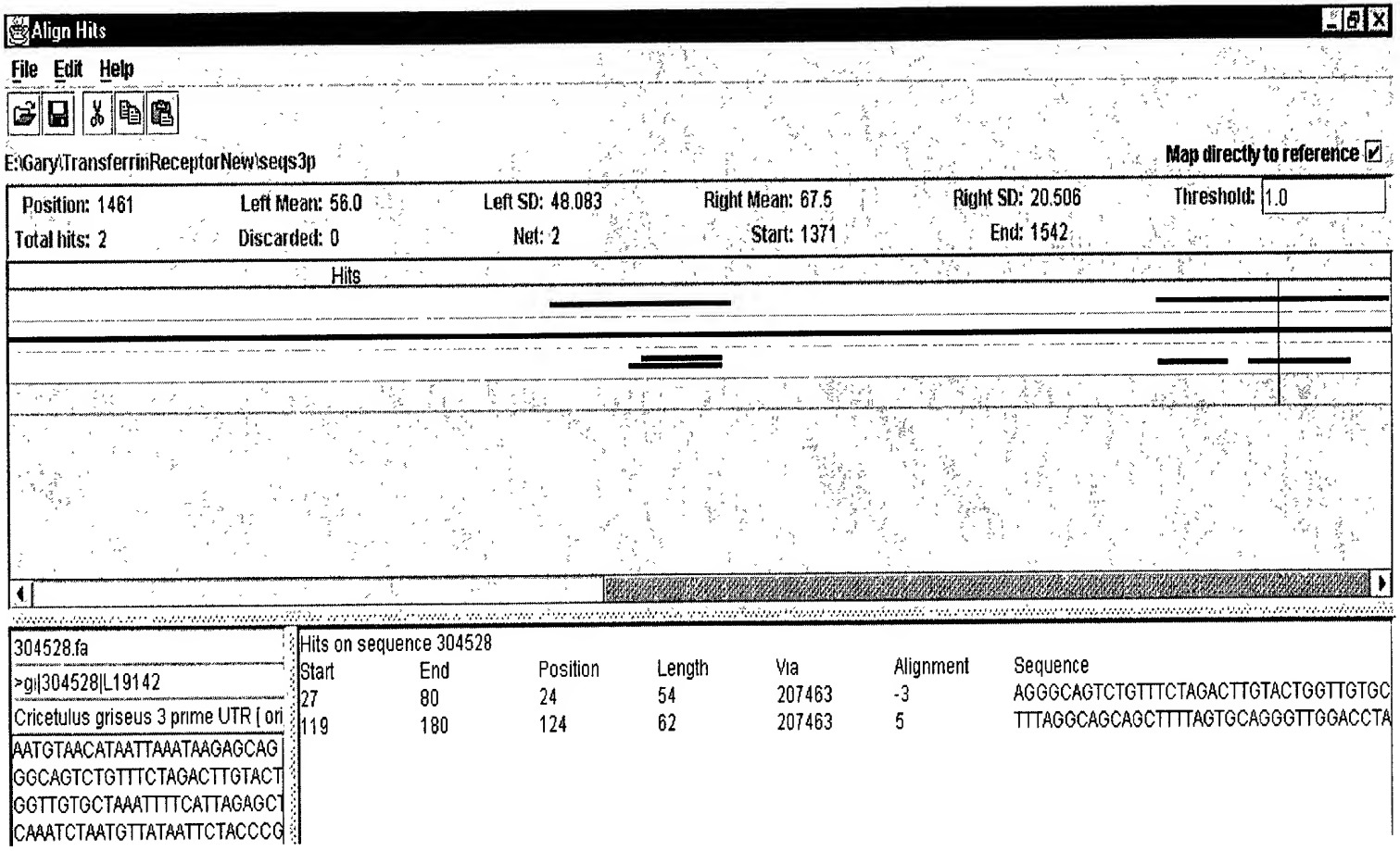

Align Hits
File Edit Help
E:\Gary\TransferrinReceptorNew\seqs3p
Map directly to reference
Position: 1461
Left Mean: 56.0
Left SD: 48.083
Right Mean: 67.5
Right SD: 20.506
Threshold: 1.0
Total hits: 2
Discarded: 0
Net: 2
Start: 1371
End: 1542
Hits
304528.fa
>gi|304528|L19142
Cricetulus griseus 3 prime UTR [ori
AATGTAACATAATTAAATAAGAGCAG
GGCAGTCTGTTTCTAGACTTGTACT
GGTTGTGCTAAATTTTCATTAGAGCT
CAAATCTAATGTTATAATTCTACCCG
Hits on sequence 304528
Start End Position Length Via Alignment Sequence
27 80 24 54 207463 -3 AGGGCAGTCTGTTTCTAGACTTGTACTGGTTGTGC
119 180 124 62 207463 5 TTTAGGCAGCAGCTTTTAGTGCAGGGTTGGACCTA

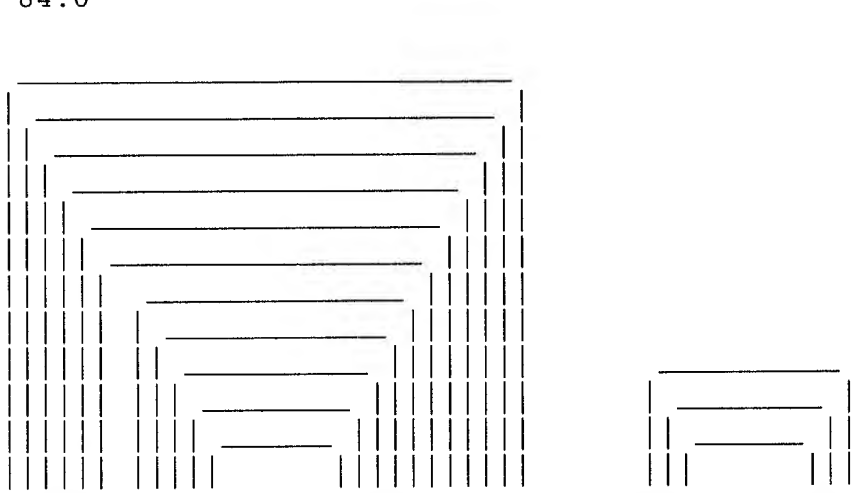
Figure 63

```
CLUSTAL W (1.74) multiple sequence alignment


gi|207463|M58040        GTACGTAATTATCGGGAACAGTGTTTCCCATAATTTTCTTCATGCGATGACATCTT
gi|37432|X01060         GTAAGTAATTATCGGGAACAGTGTTTCCCATAATTTTCTTCATGCAATGACATCTT
gi|63357|X13753         GTACGTAATTATCGGGGACAGTGTTTCCCATAATTGTTTTCATGAAATGTCATCTG
```

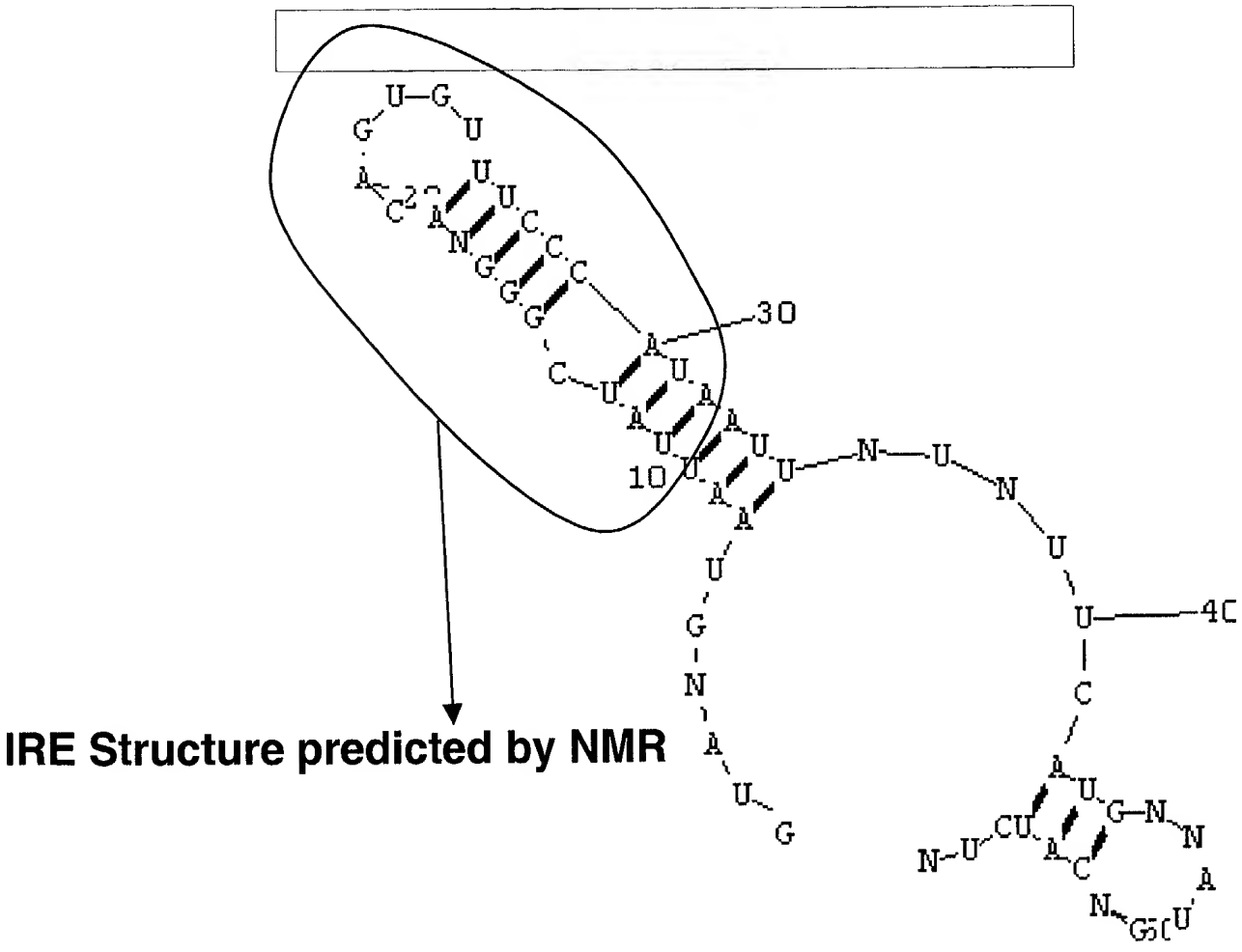
Figure 64

Score: 84.0



```
GUANGUAAUUAUCGGGNACAGUGUUUCCCAUAAUUNUNUUCAUGNNAUGNCAUCUN   Consensus
GUACGUAAUUAUCGGGAACAGUGUUUCCCAUAAUUUUCUUCAUGCGAUGACAUCUU   gi|207463|M58040
GUAAGUAAUUAUCGGGAACAGUGUUUCCCAUAAUUUUCUUCAUGCAAUGACAUCUU   gi|37432|X01060
GUACGUAAUUAUCGGGGACAGUGUUUCCCAUAAUUGUUUUCAUGAAAUGUCAUCUG   gi|63357|X13753
```

Figure 65



IRE Structure predicted by NMR

Figure 66

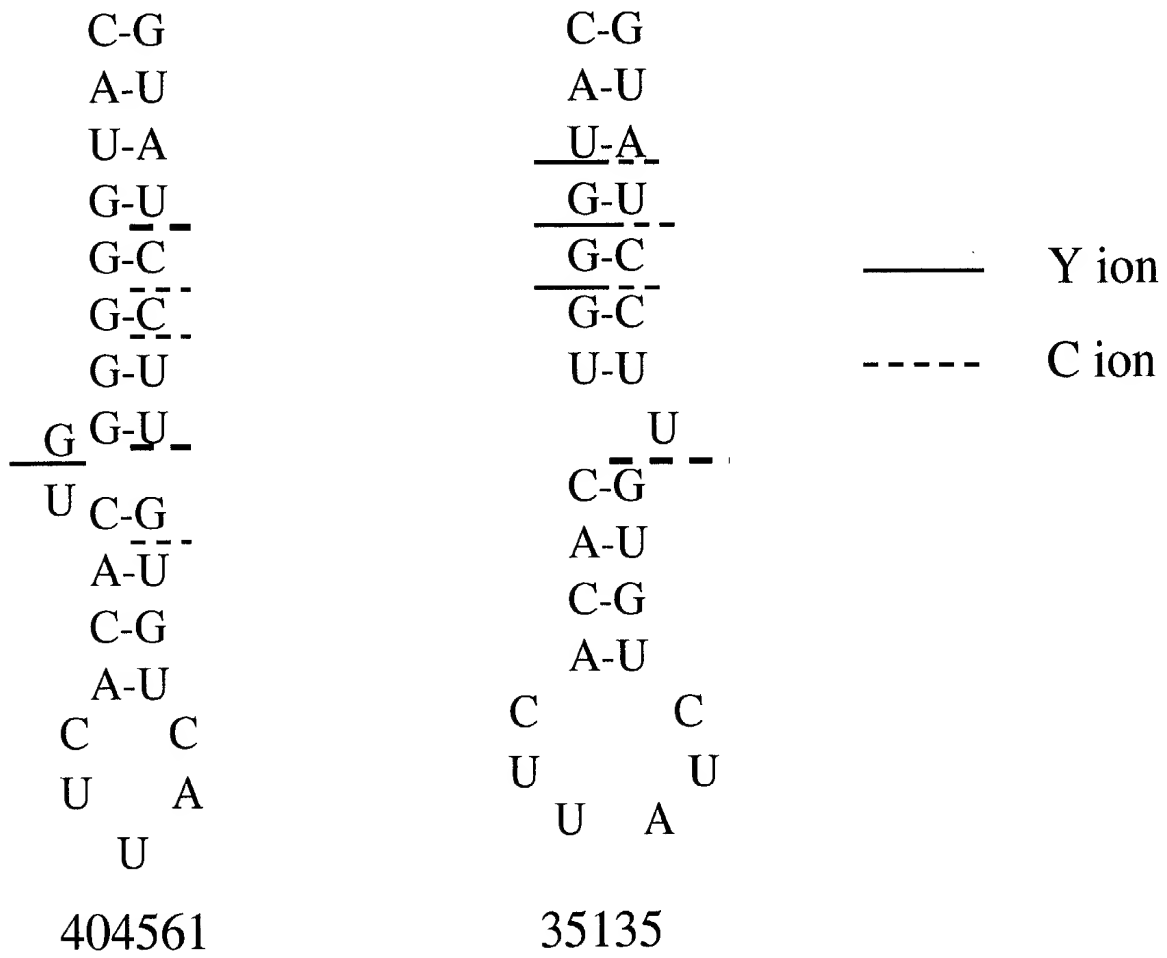
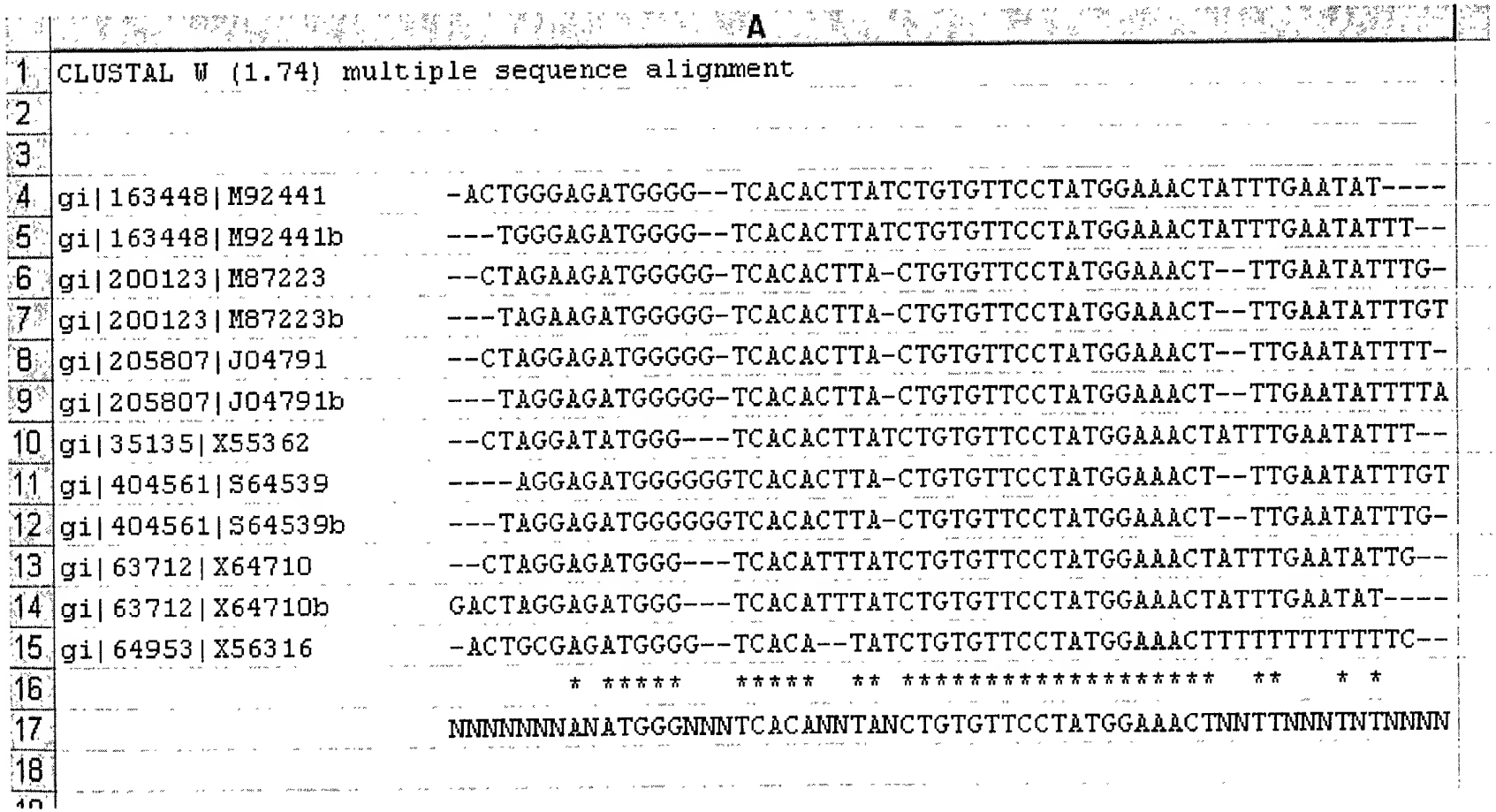


Figure 67



	A
1	CLUSTAL W (1.74) multiple sequence alignment
2	
3	
4	gi|163448|M92441 -ACTGGGAGATGGGG--TCACACTTATCTGTGTTCCTATGGAAACTATTTGAATAT----
5	gi|163448|M92441b ---TGGGAGATGGGG--TCACACTTATCTGTGTTCCTATGGAAACTATTTGAATATTT--
6	gi|200123|M87223 --CTAGAAGATGGGGG-TCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTG-
7	gi|200123|M87223b ---TAGAAGATGGGGG-TCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTGT
8	gi|205807|J04791 --CTAGGAGATGGGGG-TCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTT-
9	gi|205807|J04791b ---TAGGAGATGGGGG-TCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTTA
10	gi|35135|X55362 --CTAGGATATGGG---TCACACTTATCTGTGTTCCTATGGAAACTATTTGAATATTT--
11	gi|404561|S64539 ----AGGAGATGGGGGGTCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTGT
12	gi|404561|S64539b ---TAGGAGATGGGGGGTCACACTTA-CTGTGTTCCTATGGAAACT--TTGAATATTTG-
13	gi|63712|X64710 --CTAGGAGATGGG---TCACATTTATCTGTGTTCCTATGGAAACTATTTGAATATTG--
14	gi|63712|X64710b GACTAGGAGATGGG---TCACATTTATCTGTGTTCCTATGGAAACTATTTGAATAT----
15	gi|64953|X56316 -ACTGCGAGATGGGG--TCACA--TATCTGTGTTCCTATGGAAACTTTTTTTTTTTTC--
16	* ***** ***** ** ********** ** * *
17	NNNNNNNNANATGGGNNNTCACANNTANCTGTGTTCCTATGGAAACTNNTTNNNTNTNNNN
18	

Figure 68

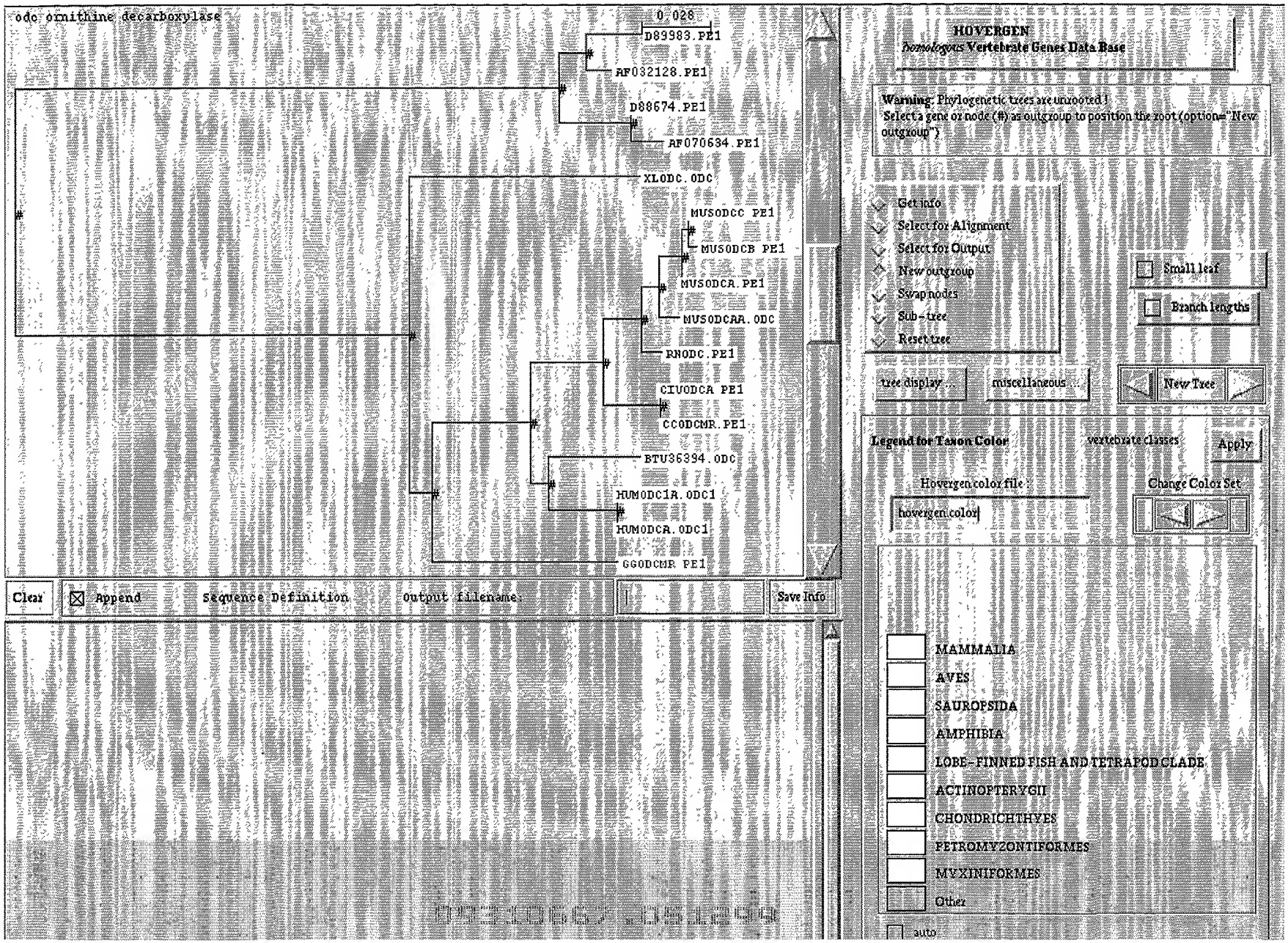

odc ornithine decarboxylase
0.028
D89983.PE1
AF032128.PE1
D88674.PE1
AF070634.PE1
XLODC.ODC
MUSODCC PE1
MUSODCB PE1
MUSODCA.PE1
MUSODCAA.ODC
RNODC.PE1
CIUODCA PE1
CCODCMR.PE1
BTU36394.ODC
HUMODC1A.ODC1
HUMODCA.ODC1
GGODCMR PE1
Clear
Append
Sequence Definition
Output filename
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted!
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display ...
miscellaneous ...
New Tree
Legend for Taxon Color
vertebrate classes
Apply
Hovergen.color.file
hovergen.color
Change Color Set
MAMMALIA
AVES
SAUROPSIDA
AMPHIBIA
LOBE-FINNED FISH AND TETRAPOD CLADE
ACTINOPTERYGII
CHONDRICHTHYES
PETROMYZONTIFORMES
MYXINIFORMES
Other
auto

Figure 69

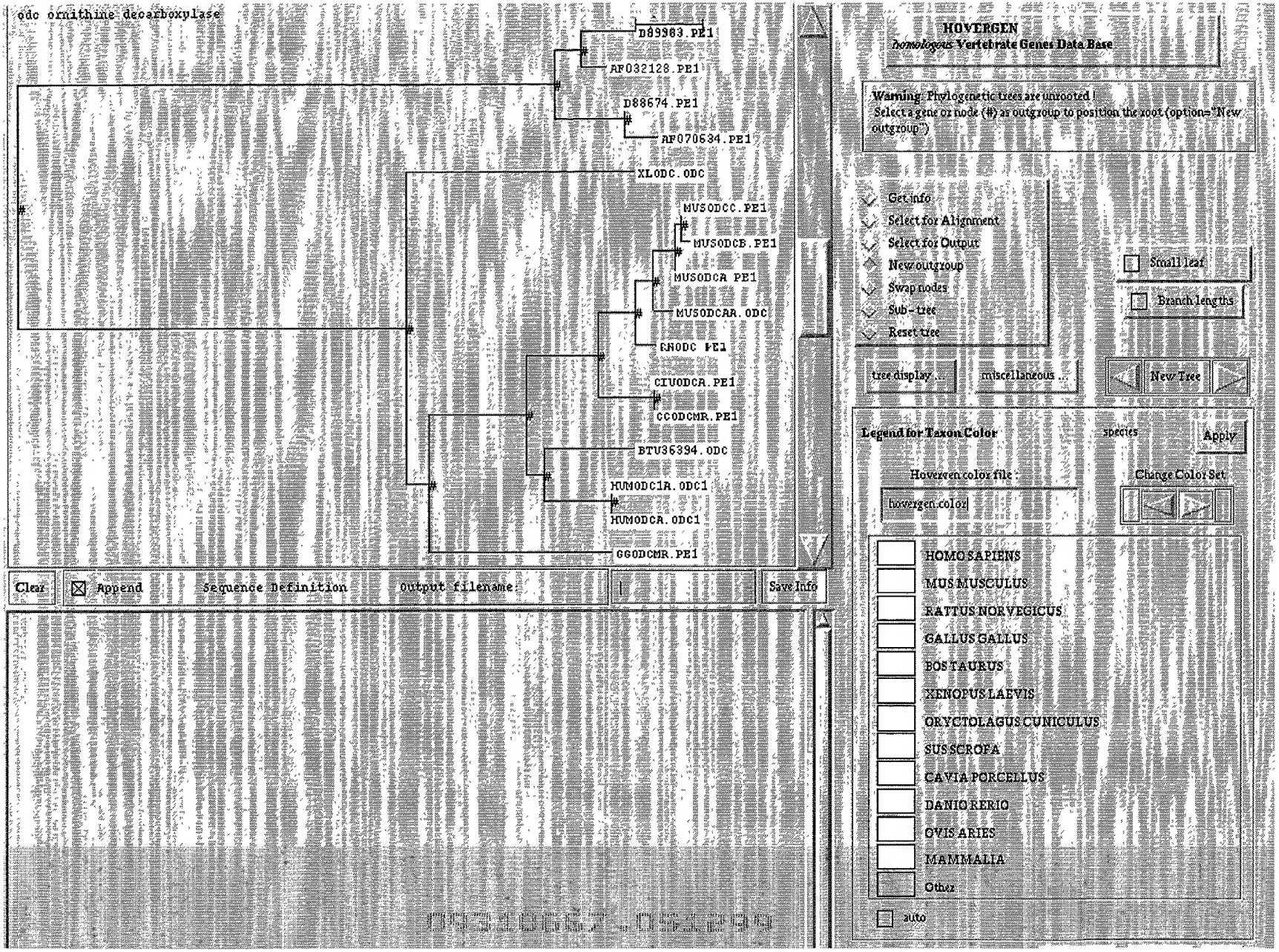

odc ornithine decarboxylase
D89983.PE1
AF032128.PE1
D88674.PE1
AF070534.PE1
XLODC.ODC
MUSODCC.PE1
MUSODCB.PE1
MUSODCA.PE1
MUSODCAA.ODC
RNODC.PE1
CIUODCA.PE1
CCODCMR.PE1
BTU36394.ODC
HUMODC1A.ODC1
HUMODCA.ODC1
GGODCMR.PE1
Clear
Append
Sequence Definition
Output filename
Save Info
HOVERGEN
homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted !
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display ...
miscellaneous ...
New Tree
Legend for Taxon Color
species
Apply
Hovergen color file :
hovergen.colour
Change Color Set
HOMO SAPIENS
MUS MUSCULUS
RATTUS NORVEGICUS
GALLUS GALLUS
BOS TAURUS
XENOPUS LAEVIS
ORYCTOLAGUS CUNICULUS
SUS SCROFA
CAVIA PORCELLUS
DANIO RERIO
OVIS ARIES
MAMMALIA
Other
auto

Figure 70

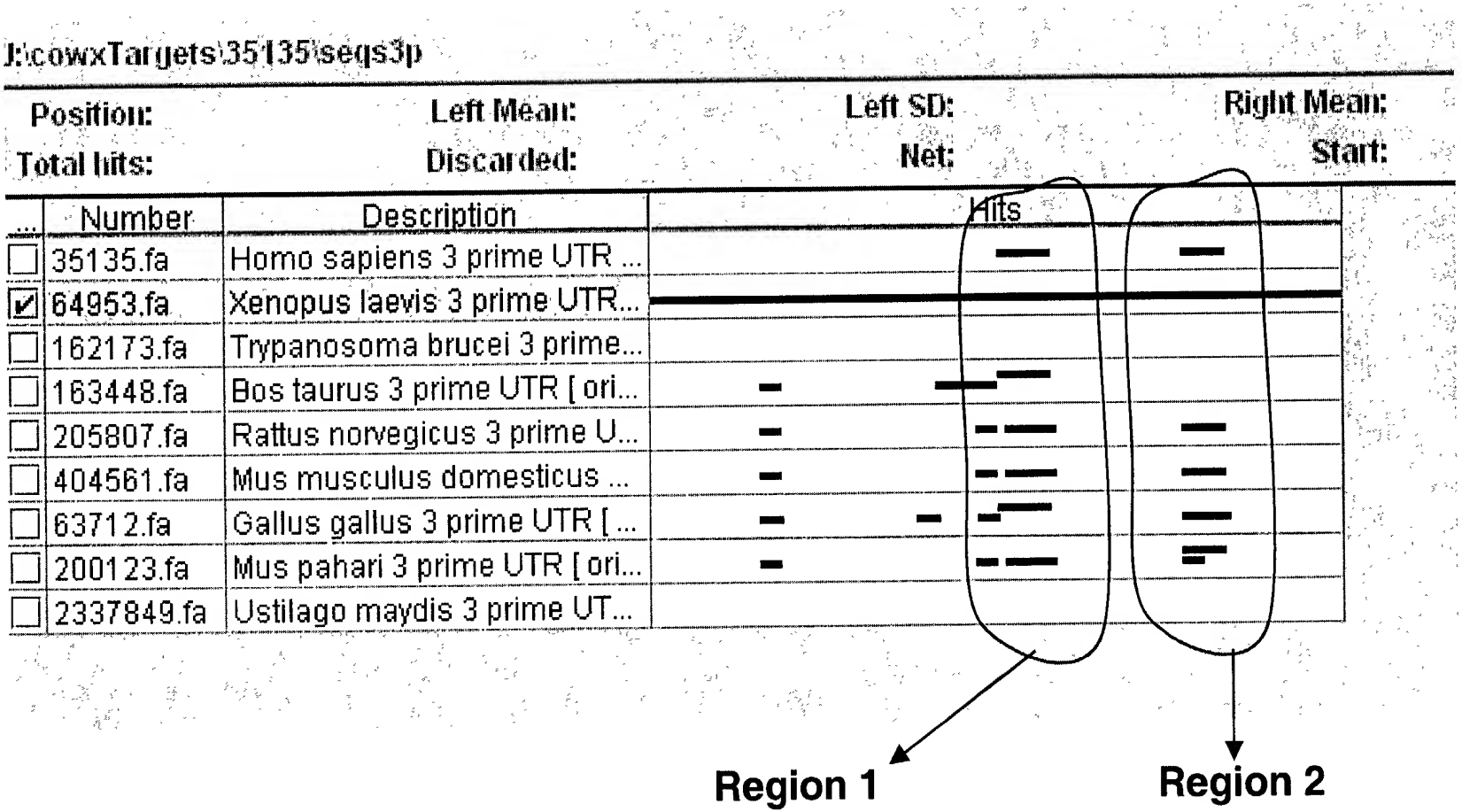


Figure 71

Figure 72

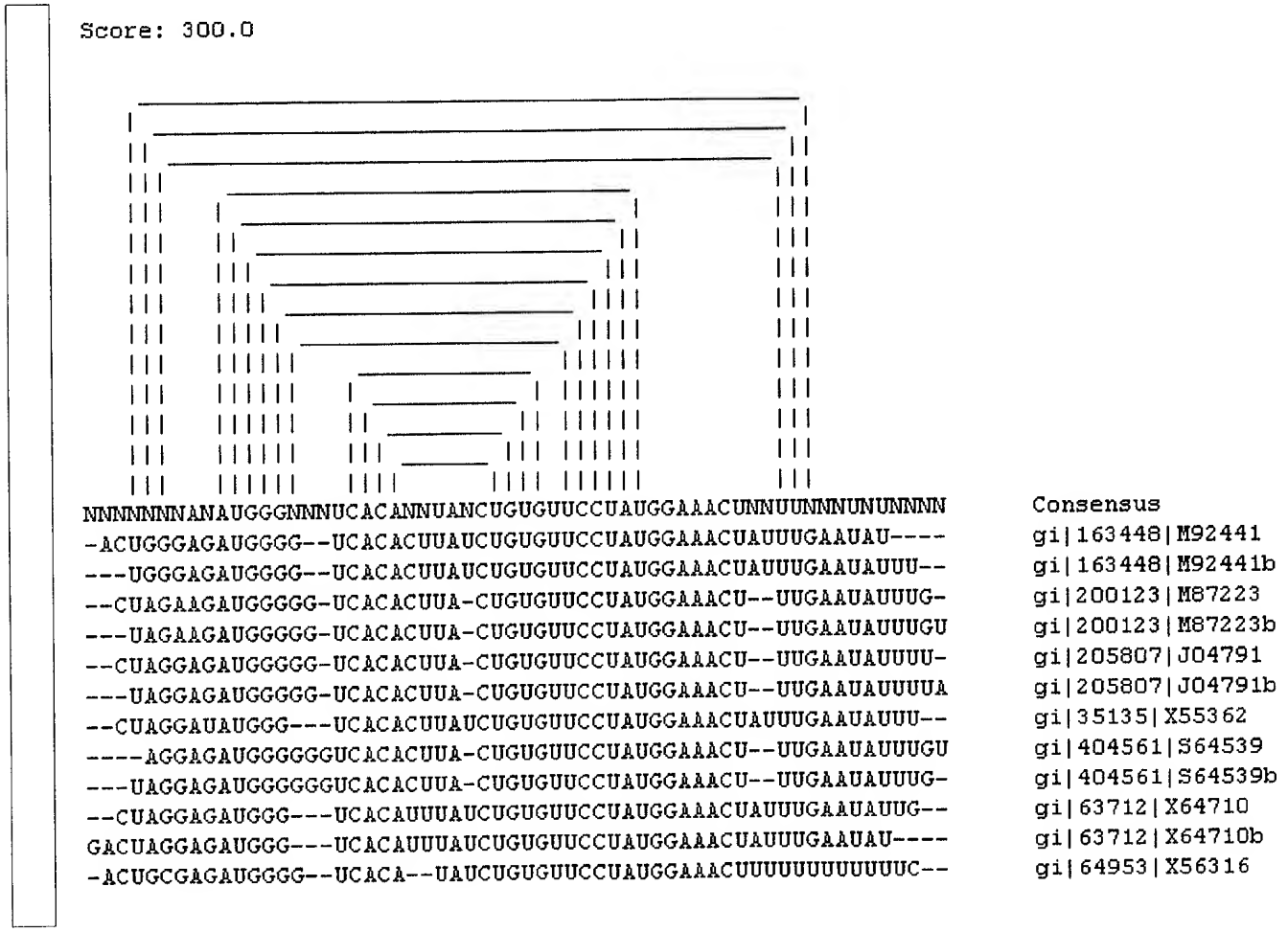


Figure 73

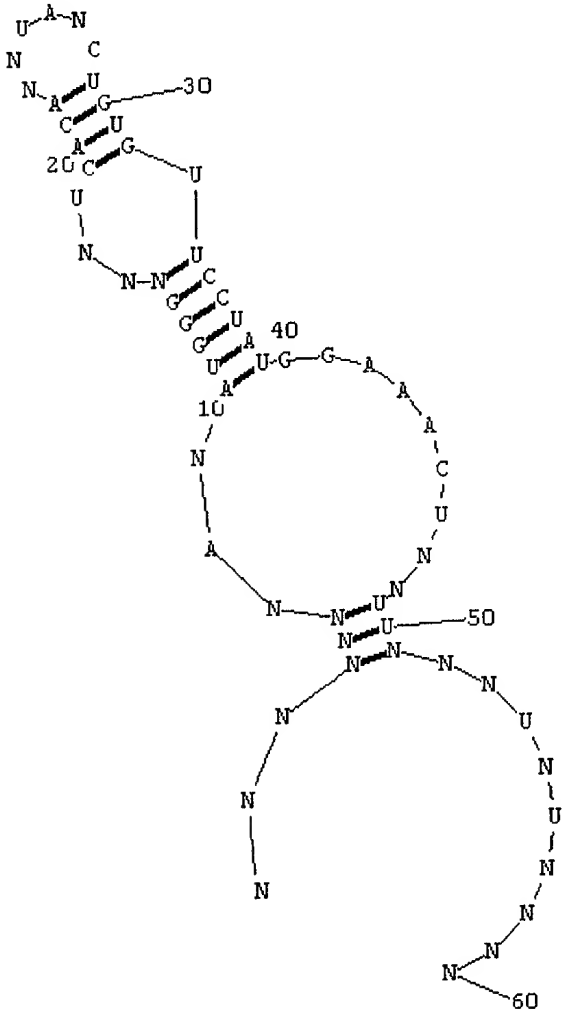


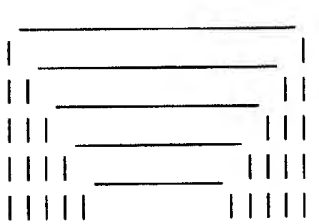
Figure 74

```
CLUSTAL W (1.74) multiple sequence alignment


gi|404561|S64539        CAAGCGTTTGTAGCTTGT
gi|63712|X64710         CAAGTATTTGTAGCTTGT
gi|64953|X56316         CAAGCATTTGTAGCTTGT
gi|35135|X55362         CAAGCATTTGTAGCTTGT
gi|200123|M87223        CAAGCATTTATAGCTTGT
gi|205807|J04791        CAAGCATTTGTAGCTTGT
                        ****  *** ********
```

Figure 75

Score: 60.0



CAAGNNUUUNUAGCUUGU	Consensus
CAAGCGUUUGUAGCUUGU	gi|404561|S64539
CAAGUAUUUGUAGCUUGU	gi|63712|X64710
CAAGCAUUUGUAGCUUGU	gi|64953|X56316
CAAGCAUUUGUAGCUUGU	gi|35135|X55362
CAAGCAUUUAUAGCUUGU	gi|200123|M87223
CAAGCAUUUGUAGCUUGU	gi|205807|J04791

Figure 76

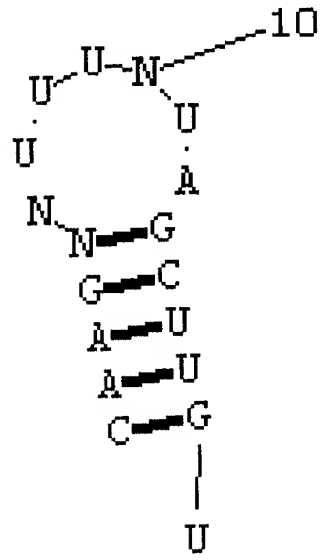


Figure 77

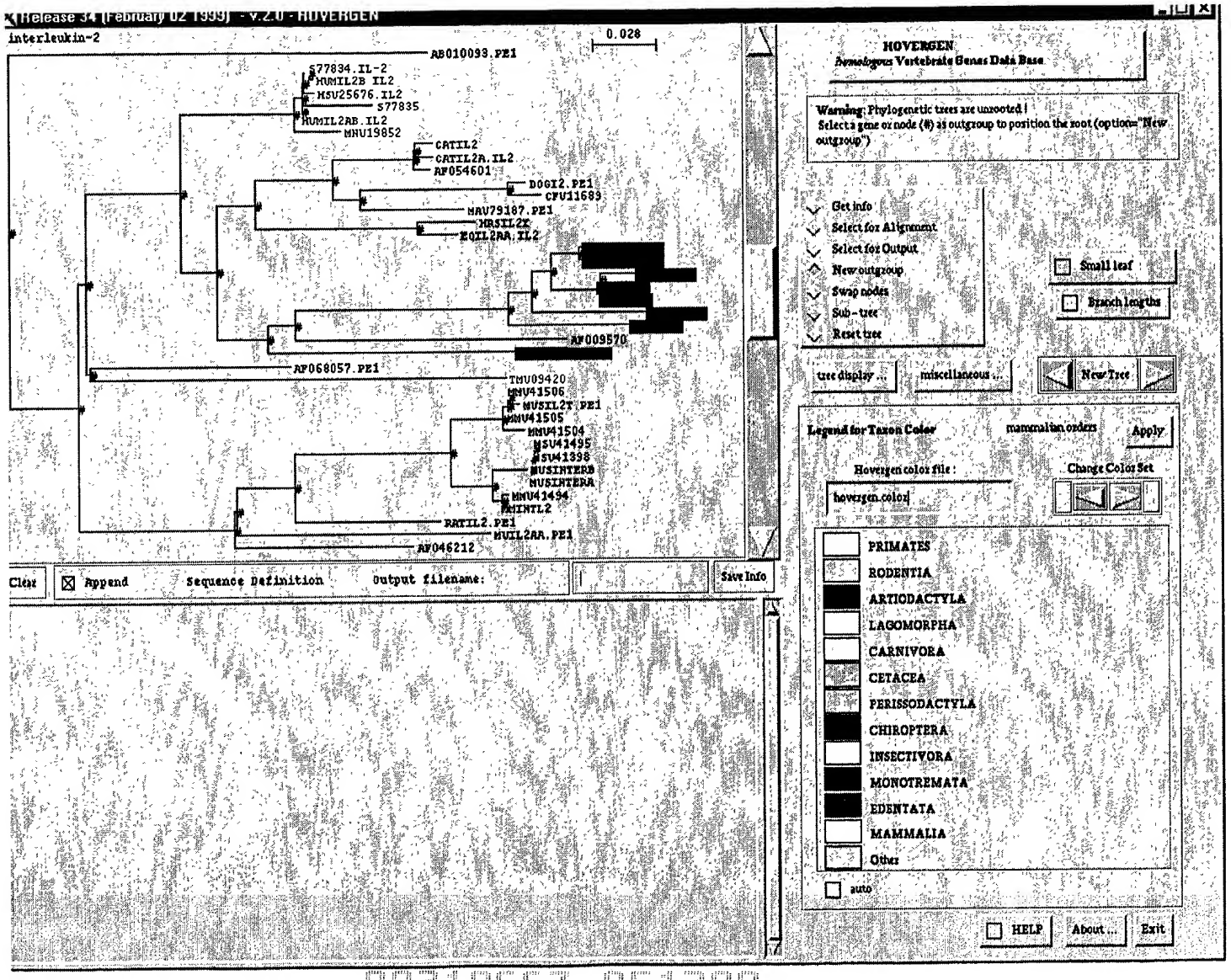

interleukin-2
0.028
AB010093.PE1
S77834.IL-2
HUMIL2B IL2
MSV25676.IL2
S77835
HUMIL2AB.IL2
MNU19852
CATIL2
CATIL2A.IL2
AF054601
DOGI2.PE1
CFU11689
MAV79187.PE1
MRSIL2X
EQIL2AA.IL2
AF009570
AF068057.PE1
TMU09420
MNU41506
MUSIL2Y.PE1
MNU41505
MNU41504
MSU41495
MSU41398
MUSINTERB
MUSINTERA
MNU41494
MINTL2
RATIL2.PE1
MUIL2AA.PE1
AF046212
Clear
Append
Sequence Definition
Output filename:
Save Info
HOVERGEN
Homologous Vertebrate Genes Data Base
Warning: Phylogenetic trees are unrooted !
Select a gene or node (#) as outgroup to position the root (option="New outgroup")
Get info
Select for Alignment
Select for Output
New outgroup
Swap nodes
Sub-tree
Reset tree
Small leaf
Branch lengths
tree display ...
miscellaneous ...
New Tree
Legend for Taxon Color
mammalian orders
Apply
Hovergen color file :
Change Color Set
hovergen.color
PRIMATES
RODENTIA
ARTIODACTYLA
LAGOMORPHA
CARNIVORA
CETACEA
PERISSODACTYLA
CHIROPTERA
INSECTIVORA
MONOTREMATA
EDENTATA
MAMMALIA
Other
auto
HELP
About ...
Exit

Figure 78

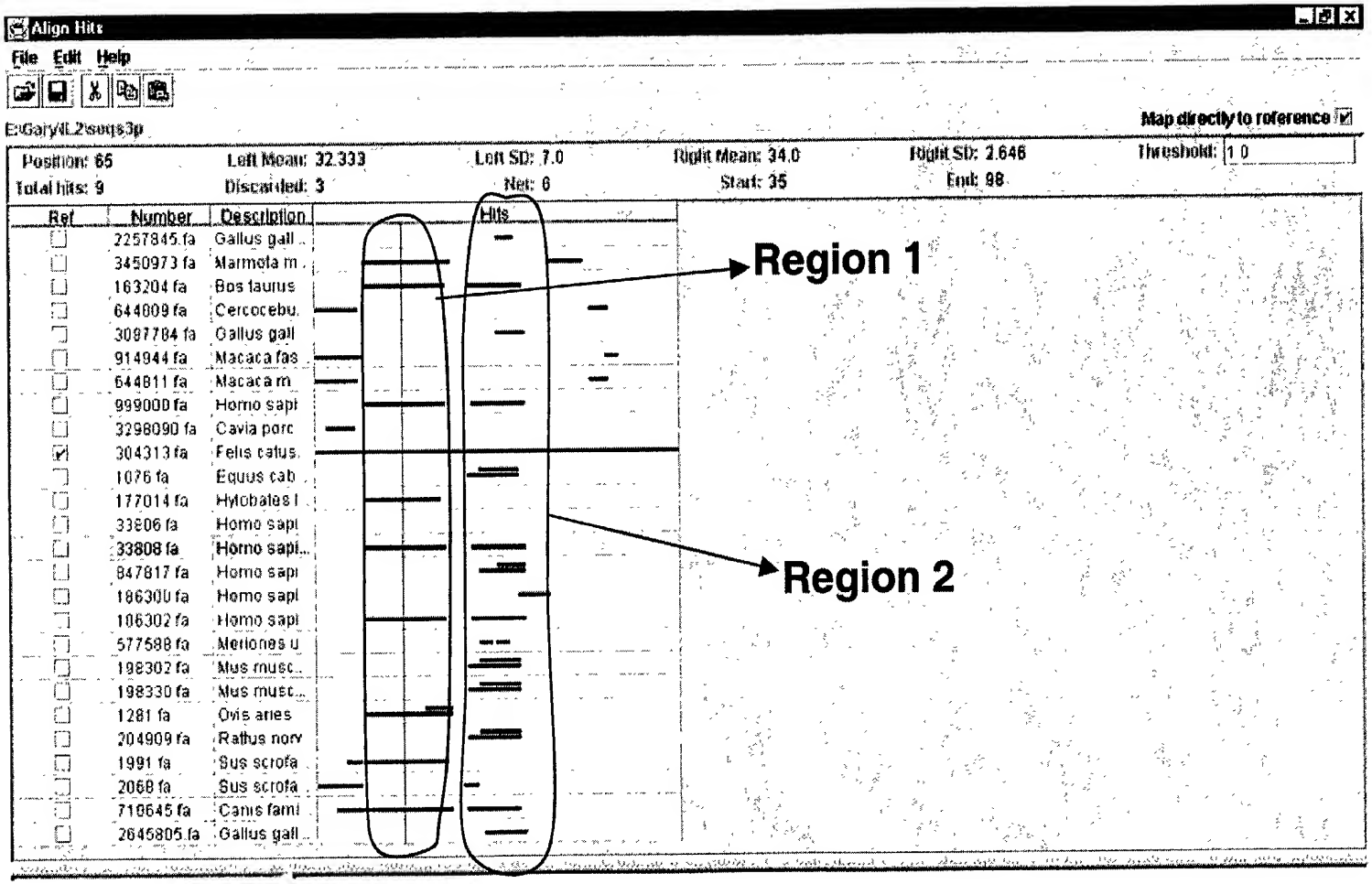


Figure 79

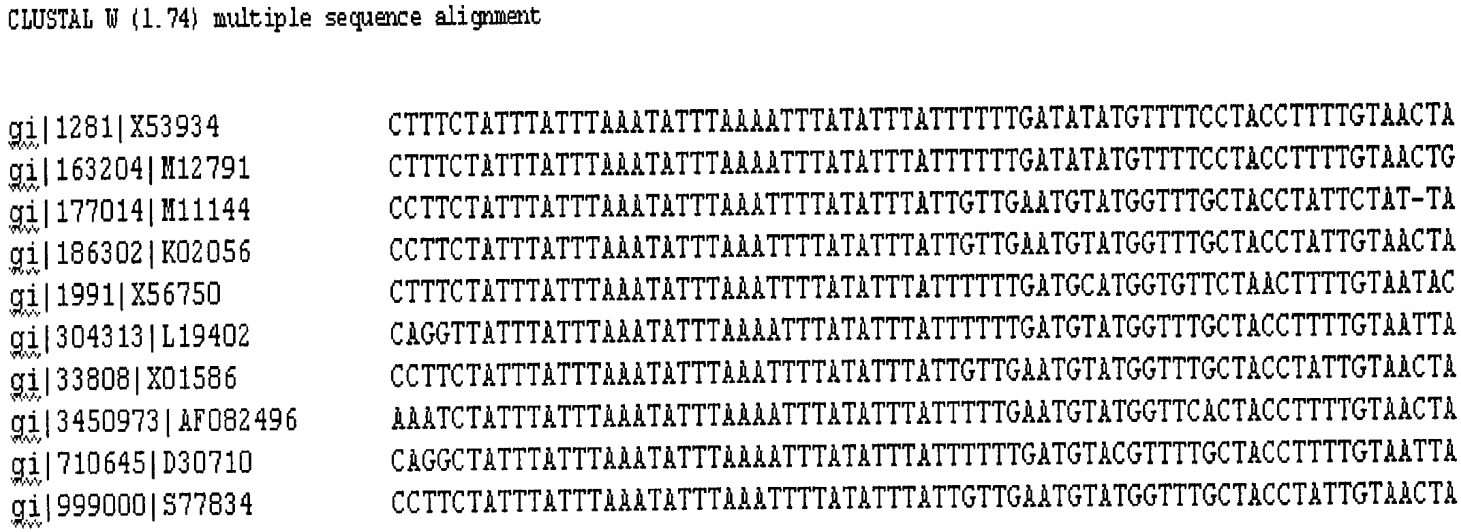
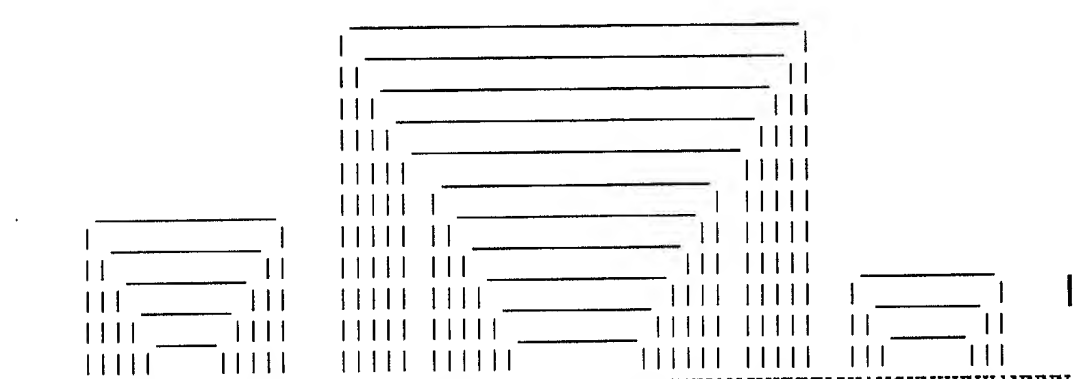


Figure 80

Score: 404.0



```
NNNNNNUAUUUAUUUAAAUAUUUAAANUUUAUAUUUAUUNUUNNAUNNANGNUNNNCUANCUNUUNUANNNN Consensus
CUUUCUAUUUAUUUAAAUAUUUAAAAUUUAUAUUUAUUUUUUGAUAUAUGUUUUCCUACCUUUUGUAACUA gi|1281|X53934
CUUUCUAUUUAUUUAAAUAUUUAAAAUUUAUAUUUAUUUUUUGAUAUAUGUUUUCCUACCUUUUGUAACUG gi|163204|M12791
CCUUCUAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUUGUUGAAUGUAUGGUUUGCUACCUAUUCUAU-UA gi|177014|M11144
CCUUCUAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUUGUUGAAUGUAUGGUUUGCUACCUAUUGUAACUA gi|186302|K02056
CUUUCUAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUUUUUUGAUGCAUGGUGUUCUAACUUUUGUAAUAC gi|1991|X56750
CAGGUUAUUUAUUUAAAUAUUUAAAAUUUAUAUUUAUUUUUUGAUGUAUGGUUUGCUACCUUUUGUAAUUA gi|304313|L19402
CCUUCUAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUUGUUGAAUGUAUGGUUUGCUACCUAUUGUAACUA gi|33808|X01586
AAAUCUAUUUAUUUAAAUAUUUAAAAUUUAUAUUUAUUUUUUGAAUGUAUGGUUCACUACCUUUUGUAACUA gi|3450973|AF082496
CAGGCUAUUUAUUUAAAUAUUUAAAAUUUAUAUUUAUUUUUUGAUGUACGUUUUGCUACCUUUUGUAAUUA gi|710645|D30710
CCUUCUAUUUAUUUAAAUAUUUAAAUUUUAUAUUUAUUGUUGAAUGUAUGGUUUGCUACCUAUUGUAACUA gi|999000|S77834
```

Figure 81

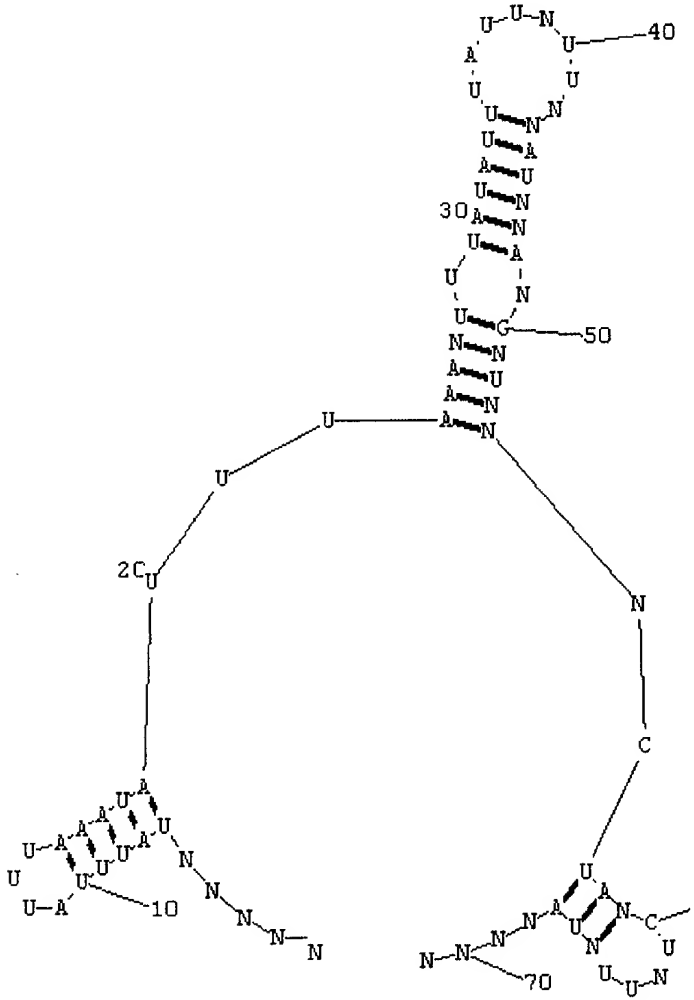


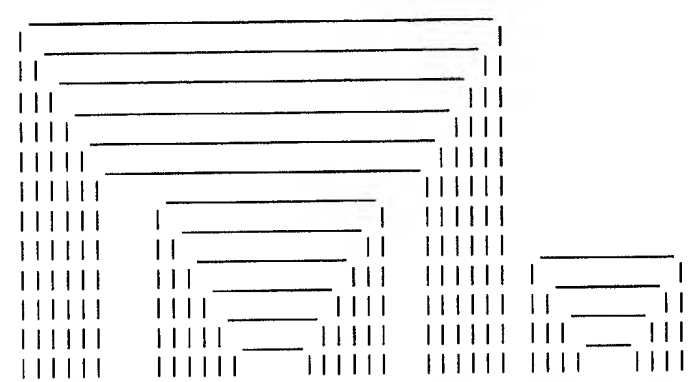
Figure 82

```
CLUSTAL W (1.74) multiple sequence alignment


gi|1076|X69393                TGATAAATATG--GATCTTTTAAGATTCTTTTTGTAAG-CCCXXXXXXX-
gi|163204|M12791              TGATAAATATG--GATCTTTTAAGATTCTTTTTGTAAG-CCCTACGGGC-
gi|186300|M22005              XXXXXXXXXXX--XXXXXXXXXXXXXXXXXXXXXXXXTCCACCCTGA--
gi|186302|K02056              CTATAAATATG--GATCTTTTATGATTCTTTTTGTAAG-CCCTAGGGGC-
gi|198302|K02797              TGATAAATATG--GATCTTTAAAGATTCTTTTTGTAAGCCCCAAGGGC--
gi|198330|K02292              TGATAAATATG--GATCTTTAAAGATTCTTTTTGTAAGCCCCAAGGGC--
gi|204909|M22899              TGATAAATATG--GATCTTTAAAGATTCTTTTTGTAAGCCCCAAGGGC--
gi|2068|X58428                TGTTAAATGTATCAGGCXXXXXXXXXXXXXXXXXXXXX-XXXXXXXX---
gi|2257845|AF000631           -GTCTGTTCTAA-GATCATATTTGATCCTTTCTGTAA--CCTACGGGCTC
gi|2645805|AF033563           -GTCTGTTCTAA-GATCATATTTGATCCTTTCTGTAAGCCCTACGGGC--
gi|304313|L19402              TGATAAACATG--GATCTTTTAAGATTCTTTTTGTAAG-CCCTAGGCGT-
gi|3087784|AJ224516           -GTCTGTTCTAA-GATCATATTTGATCCTTTCTGTAAGCCCTACGGGC--
gi|33808|X01586               CTATAAATATG--GATCTTTTATGATTCTTTTTGTAAG-CCCTAGGGGC-
gi|577588|X68779              CGATAAATATG--GATCTTTAGAGATTCTTTTTTGTA--CCCAAGGGCTC
gi|710645|D30710              TGATAAATATG--GATCTTTTAAGATTCTTTTTGTAAG-CCCXXXXXXX-
gi|847817|U25676              CTATAAATATG--GATCTTTTATGATTCTTTTTGTAAG-CCCTAGGGGC-
gi|999000|S77834              CTATAAATATG--GATCTTTTATGATTCTTTTTGTAAG-CCCTAGGGGC-
```

Figure 83

Score: 412.0



```
NNNUNNNNNNNNGAUCNUNNNNGAUNCUUUNUNNNANNCCNNNNNNNN  Consensus
UGAUAAAAUAUGGAUCUUUUAAGAUUCUUUUUGUAAGCCCXXXXXXX  gi|1076|X69393
UGAUAAAAUAUGGAUCUUUUAAGAUUCUUUUUGUAAGCCCUACGGGC  gi|163204|M12791
CUAUAAAAUAUGGAUCUUUUAUGAUUCUUUUUGUAAGCCCUAGGGGC  gi|186302|K02056
UGAUAAAAUAUGGAUCUUUAAAGAUUCUUUUUGUAAGCCCCAAGGGC  gi|198302|K02797
UGAUAAAAUAUGGAUCUUUAAAGAUUCUUUUUGUAAGCCCCAAGGGC  gi|198330|K02292
UGAUAAAAUAUGGAUCUUUAAAGAUUCUUUUUGUAAGCCCCAAGGGC  gi|204909|M22899
GUCUGUUCUAAGAUCAUAUUUGAUCCUUUCUGUAA--CCUACGGGC  gi|2257845|AF000631
GUCUGUUCUAAGAUCAUAUUUGAUCCUUUCUGUAAGCCCUACGGGC  gi|2645805|AF033563
UGAUAAACAUGGAUCUUUUAAGAUUCUUUUUGUAAGCCCUAGGCGU  gi|304313|L19402
GUCUGUUCUAAGAUCAUAUUUGAUCCUUUCUGUAAGCCCUACGGGC  gi|3087784|AJ224516
CUAUAAAAUAUGGAUCUUUUAUGAUUCUUUUUGUAAGCCCUAGGGGC  gi|33808|X01586
CGAUAAAAUAUGGAUCUUUAGAGAUUCUUUUUUGUA--CCCAAGGGC  gi|577588|X68779
UGAUAAAAUAUGGAUCUUUUAAGAUUCUUUUUGUAAGCCCXXXXXXX  gi|710645|D30710
CUAUAAAAUAUGGAUCUUUUAUGAUUCUUUUUGUAAGCCCUAGGGGC  gi|847817|U25676
CUAUAAAAUAUGGAUCUUUUAUGAUUCUUUUUGUAAGCCCUAGGGGC  gi|999000|S77834
```

Figure 84

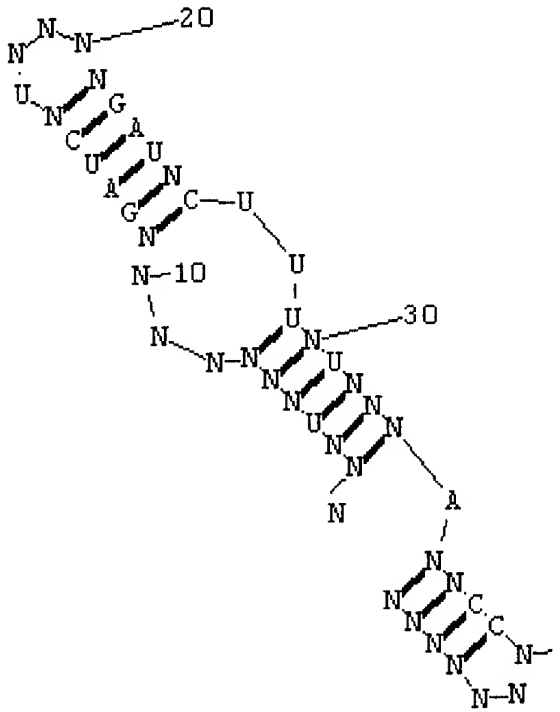


Figure 85

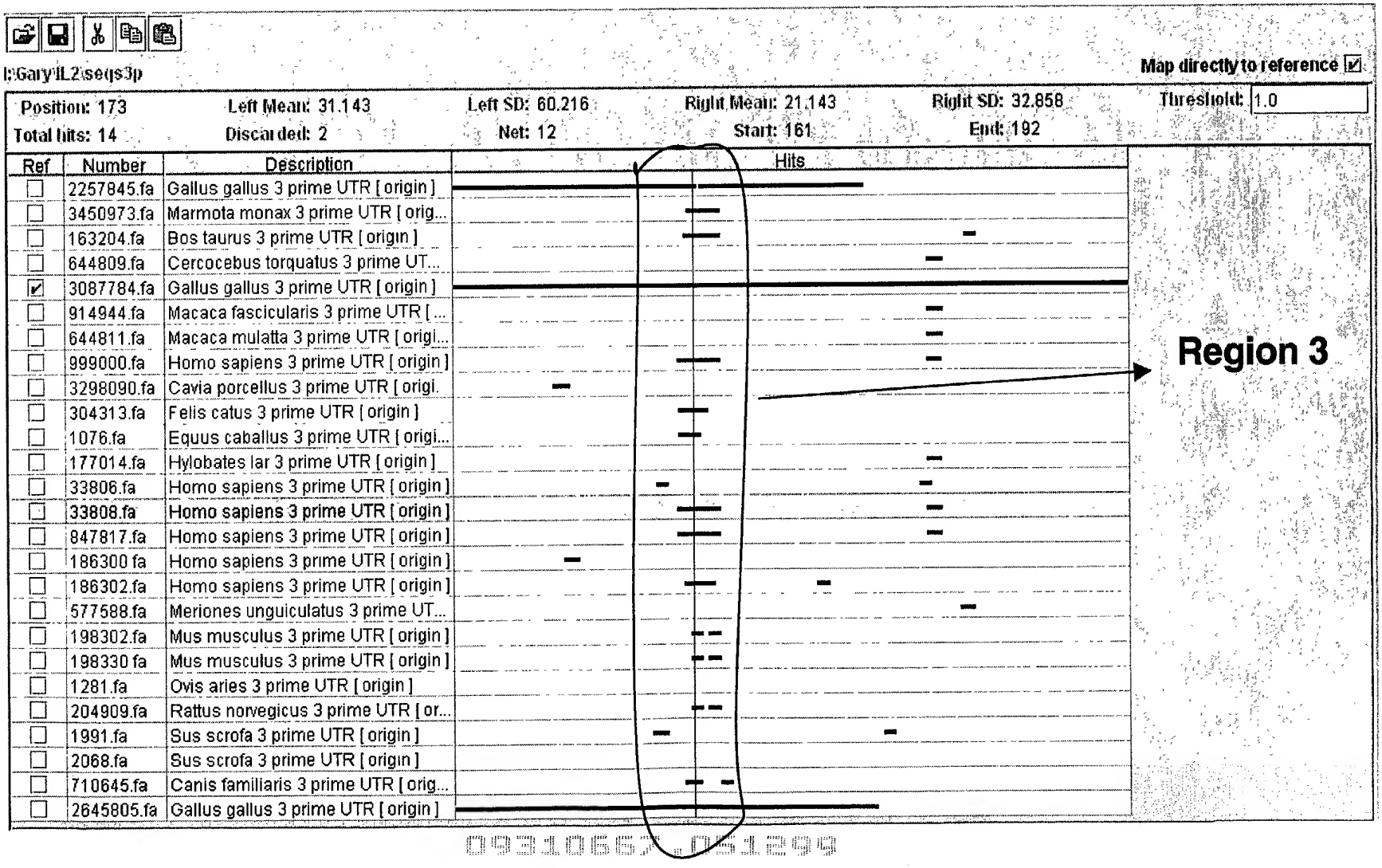

I:\Gary\L2\seqs3p
Map directly to reference
Position: 173
Left Mean: 31.143
Left SD: 60.216
Right Mean: 21.143
Right SD: 32.858
Threshold: 1.0
Total hits: 14
Discarded: 2
Net: 12
Start: 161
End: 192
Ref
Number
Description
Hits
2257845.fa Gallus gallus 3 prime UTR [origin]
3450973.fa Marmota monax 3 prime UTR [orig...
163204.fa Bos taurus 3 prime UTR [origin]
644809.fa Cercocebus torquatus 3 prime UT...
3087784.fa Gallus gallus 3 prime UTR [origin]
914944.fa Macaca fascicularis 3 prime UTR [...
644811.fa Macaca mulatta 3 prime UTR [origi...
999000.fa Homo sapiens 3 prime UTR [origin]
3298090.fa Cavia porcellus 3 prime UTR [origi.
304313.fa Felis catus 3 prime UTR [origin]
1076.fa Equus caballus 3 prime UTR [origi...
177014.fa Hylobates lar 3 prime UTR [origin]
33806.fa Homo sapiens 3 prime UTR [origin]
33808.fa Homo sapiens 3 prime UTR [origin]
847817.fa Homo sapiens 3 prime UTR [origin]
186300.fa Homo sapiens 3 prime UTR [origin]
186302.fa Homo sapiens 3 prime UTR [origin]
577588.fa Meriones unguiculatus 3 prime UT...
198302.fa Mus musculus 3 prime UTR [origin]
198330.fa Mus musculus 3 prime UTR [origin]
1281.fa Ovis aries 3 prime UTR [origin]
204909.fa Rattus norvegicus 3 prime UTR [or...
1991.fa Sus scrofa 3 prime UTR [origin]
2068.fa Sus scrofa 3 prime UTR [origin]
710645.fa Canis familiaris 3 prime UTR [orig...
2645805.fa Gallus gallus 3 prime UTR [origin]
Region 3

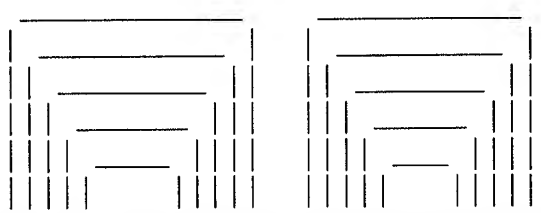
Figure 86

```
CLUSTAL W (1.74) multiple sequence alignment


gi|163204|M12791          -TTTTAAGATTCTTTTTGTAAGCCCTACGGGCT-TAAAAATTCA
gi|304313|L19402          -TTTTAAGATTCTTTTTGTAAGCCCTAGGCGTGCTAAAAACTC-
gi|198302|K02797          -TTTAAAGATTCTTTTTGTAAGCCCCAAGGGCTCAAAAATGTT-
gi|204909|M22899          -TTTAAAGATTCTTTTTGTAAGCCCCAAGGGCTCAAAAATGTT-
gi|3450973|AF082496       CTTTTAAGATTCTTTT-GTAAGCCCTAAGGGCTCTAAAATGGT-
gi|33808|X01586           -TTTTATGATTCTTTTTGTAAGCCCTAGGGGCTCTAAAATGGT-
gi|3087784|AJ224516       -ATATTTGATCCTTTCTGTAAGCCCTACGGGCTCAAAATGTAC-
                            *    *** ****  ******** * * *    ***
```

Figure 87

Score: 165.0



```
NNUNNNNGAUNCUUUNNGUAAGCCCNANGNGNNNNAAANNNNNN    Consensus
-UUUUAAGAUUCUUUUUGUAAGCCCUACGGGCU-UAAAAAUUCA    gi|163204|M12791
-UUUUAAGAUUCUUUUUGUAAGCCCUAGGCGUGCUAAAAACUC-    gi|304313|L19402
-UUUAAAGAUUCUUUUUGUAAGCCCCAAGGGCUCAAAAAUGUU-    gi|198302|K02797
-UUUAAAGAUUCUUUUUGUAAGCCCCAAGGGCUCAAAAAUGUU-    gi|204909|M22899
CUUUUAAGAUUCUUUU-GUAAGCCCUAAGGGCUCUAAAAUGGU-    gi|3450973|AF082496
-UUUUAUGAUUCUUUUUGUAAGCCCUAGGGGCUCUAAAAUGGU-    gi|33808|X01586
-AUAUUUGAUCCUUUCUGUAAGCCCUACGGGCUCAAAAUGUAC-    gi|3087784|AJ224516
```

Figure 88

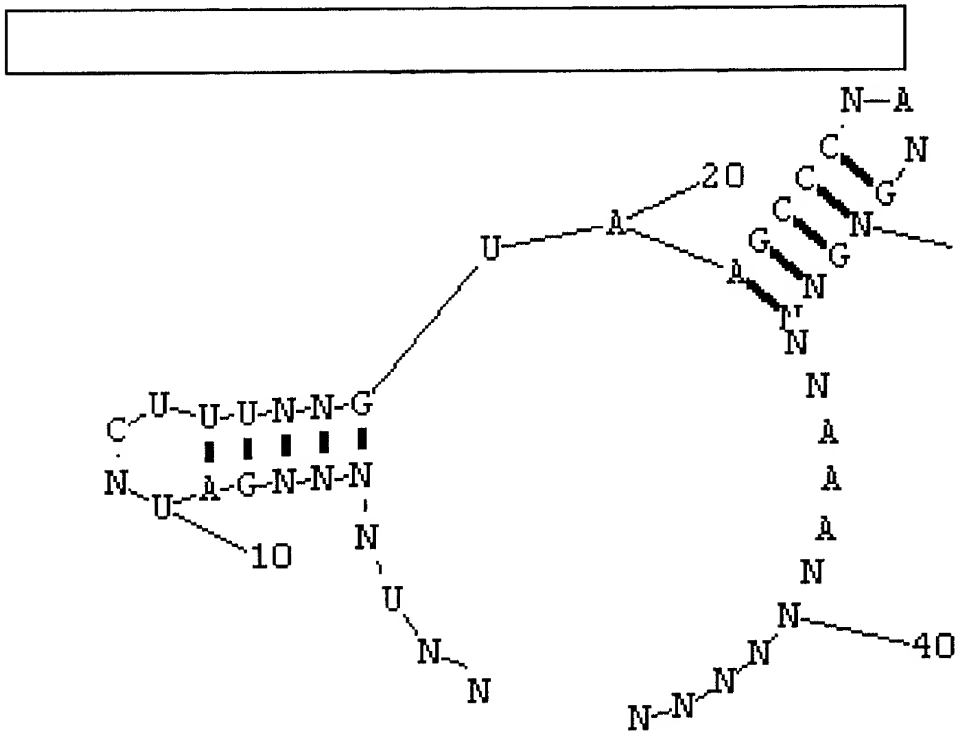


Figure 89

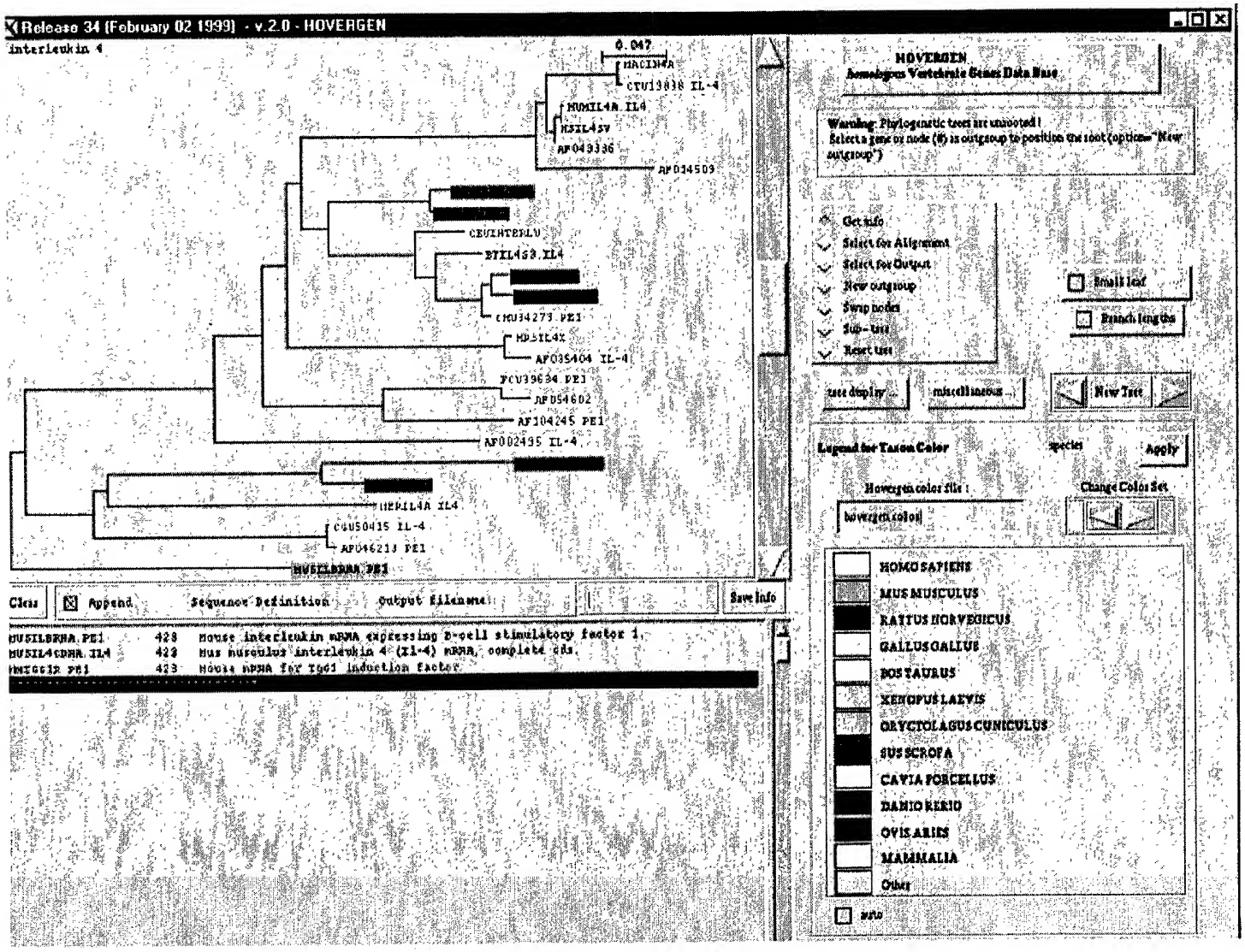


Figure 90

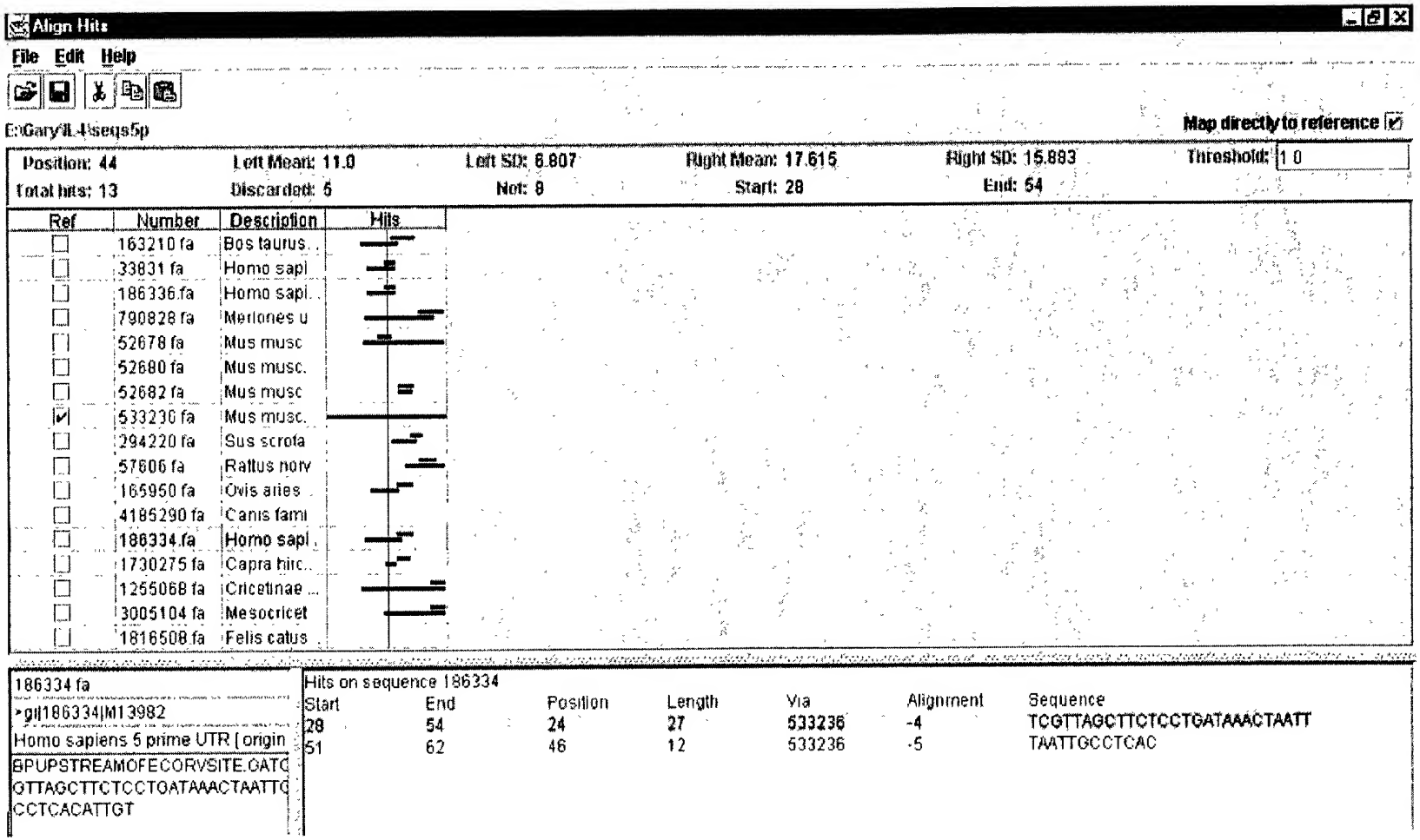

Align Hits
File Edit Help
E:\Gary\L4\seqs5p
Map directly to reference
Position: 44
Left Mean: 11.0
Left SD: 6.807
Right Mean: 17.615
Right SD: 15.893
Threshold: 1.0
Total hits: 13
Discarded: 5
Not: 8
Start: 28
End: 54
Ref
Number
Description
Hits
163210.fa Bos taurus..
33831.fa Homo sapi
186336.fa Homo sapi..
790828.fa Merlones u
52678.fa Mus musc
52680.fa Mus musc.
52682.fa Mus musc
533230.fa Mus musc.
294220.fa Sus scrota
57806.fa Rattus norv
165950.fa Ovis aries
4186290.fa Canis fami
186334.fa Homo sapi..
1730275.fa Capra hirc..
1255068.fa Cricetinae ...
3005104.fa Mesocricet
1816508.fa Felis catus
186334.fa
>gi|186334|M13982
Homo sapiens 5 prime UTR [origin
8PUPSTREAMOFECORVSITE.GATC
GTTAGCTTCTCCTGATAAACTAATTC
CCTCACATTGT
Hits on sequence 186334
Start End Position Length Via Alignment Sequence
28 54 24 27 533236 -4 TCGTTAGCTTCTCCTGATAAACTAATT
51 62 46 12 533236 -5 TAATTGCCTCAC

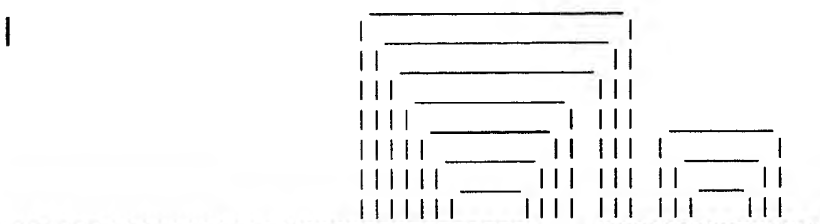
Figure 91

```
CLUSTAL W (1.74) multiple sequence alignment

gi|1255068|U50415        XXXTGCATTGTTAGCGTCTCTTGATAACCTTAAT-TGTCTCCTATCACTGA---
gi|163210|M77120         -XXXXTGCATTGTTAGCGTCTCCTGGTAAACTAATTGTCTCACATTGTCAGT--
gi|165950|M96845         -XXXXXXXXXXXXTAGCTTCTCCTGATAATCTAATTGCCTCACACTGTCAGT--
gi|1730275|U34273        -XXXXXXXXXXXXTAGCTTCTCCTGATAAACTAATTGTCTCACACTGTCAGT--
gi|186334|M13982         --------GATCGTTAGCTTCTCCTGATAAACTAATTGCCTCACATTGTXXX---
gi|186336|M23442         -GCATTGCATCGTTAGCTTCTCCTGATAAACTAATTGCCTCACATTGTCACT--
gi|294220|L12991         ---XXXXXXXXXXXXXXXXXXXXGGTAAACTAATTGTCTCACATCGTCAGTGC
gi|3005104|AF046213      XXXXXXXXXXXXXXXXXXXXXTTGATAACCTTAATTGTCTCCTATCACTGA---
gi|33831|X06750          -GCATTGCATCGTTAGCTTCTCCTGATAAACTAATTGCCTCACATTGTCACT--
gi|52678|X05064          GCATTGCATTGTTAGCATCTCTTGATAAACTTAATTGTCTCTCGTCACTGA---
gi|52678|X05064b         -GTGTGCGCAAGGCAGACTTTCTTGATA-TTACTCTGTCTTTCCCCAGGGCGA-
gi|52682|X05253          XXXXXXXXCCCAGGCCCGACAGCGAGACCCAAATCTGTCTCACAATGAAAC---
gi|533236|M25892         GGGGGGATTTGTTAGCATCTCTTGATAAACTTAATTGTCTCTCGTCACTGA---
gi|790828|L37779         XXXXXXXATTGTTGGCATCTCTTGACAAACTTAATTGTCTCACATCCCTGA---
```

Figure 92

Score: 172.0



```
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNUGNCUNNNNNNNNNNNNNNN   Consensus
XXXUGCAUUGUUAGCGUCUCUUGAUAACCUUAAU-UGUCUCCUAUCACUGA---         gi|1255068|U50415
-XXXXUGCAUUGUUAGCGUCUCCUGGUAAACUAAUUGUCUCACAUUGUCAGU--         gi|163210|M77120
-XXXXXXXXXXXXUAGCUUCUCCUGAUAAUCUAAUUGCCUCACACUGUCAGU--         gi|165950|M96845
-XXXXXXXXXXXXUAGCUUCUCCUGAUAAACUAAUUGUCUCACACUGUCAGU--         gi|1730275|U34273
-------GAUCGUUAGCUUCUCCUGAUAAACUAAUUGCCUCACAUUGUXXX---         gi|186334|M13982
-GCAUUGCAUCGUUAGCUUCUCCUGAUAAACUAAUUGCCUCACAUUGUCACU--         gi|186336|M23442
---XXXXXXXXXXXXXXXXXXXXGGUAAACUAAUUGUCUCACAUCGUCAGUGC         gi|294220|L12991
XXXXXXXXXXXXXXXXXXXXXUUGAUAACCUUAAUUGUCUCCUAUCACUGA---         gi|3005104|AF046213
-GCAUUGCAUCGUUAGCUUCUCCUGAUAAACUAAUUGCCUCACAUUGUCACU--         gi|33831|X06750
GCAUUGCAUUGUUAGCAUCUCUUGAUAAACUUAAUUGUCUCUCGUCACUGA---         gi|52678|X05064
-GUGUGCGCAAGGCAGACUUUCUUGAUA-UUACUCUGUCUUUCCCCAGGGCGA-         gi|52678|X05064b
XXXXXXXXCCCAGGCCCGACAGCGAGACCCAAAUCUGUCUCACAAUGAAAC---         gi|52682|X05253
GGGGGGAUUUGUUAGCAUCUCUUGAUAAACUUAAUUGUCUCUCGUCACUGA---         gi|533236|M25892
XXXXXXXAUUGUUGGCAUCUCUUGACAAACUUAAUUGUCUCACAUCCCUGA---         gi|790828|L37779
```

Figure 93

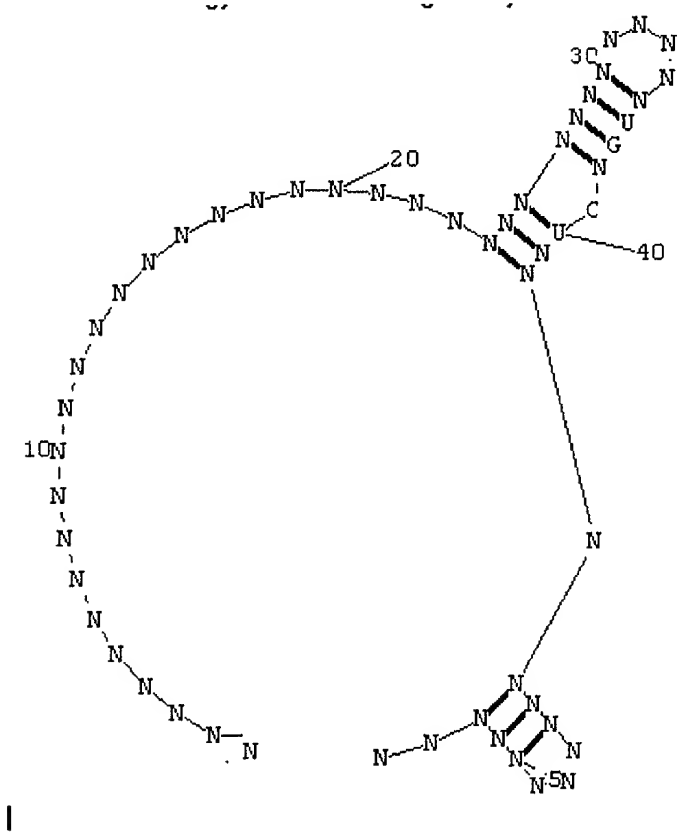


Figure 94

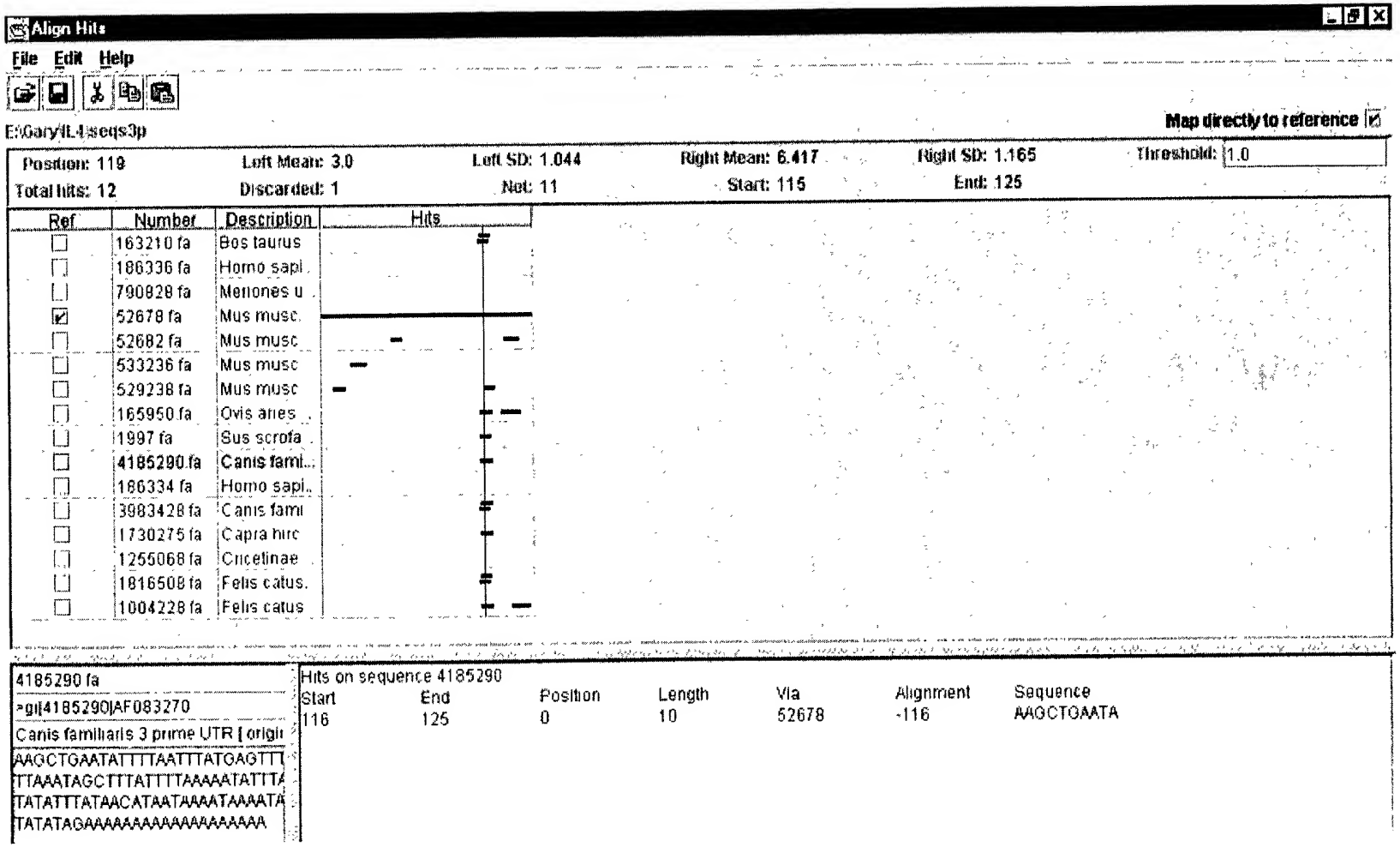


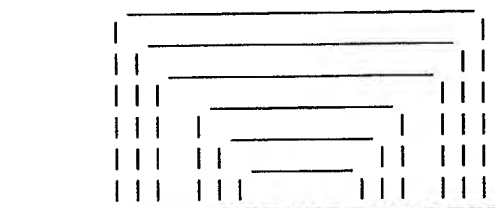
Figure 95

```
CLUSTAL W (1.74) multiple sequence alignment


gi|1004228|X87408          TTATAATTATTATAAAATAAGTATATGT-
gi|1004228|X87408b         XXXXXAAGCTGAATATCTTAATTTATGAG
gi|163210|M77120           ATGTAGAGCTGAAAAAAXXXXXXXXXXXX
gi|163210|M77120b          XXXXXAAGCTGAATATTTTAATTTATGAG
gi|165950|M96845           AATTTATGCTTTTTAATAGCTTATAT---
gi|165950|M96845b          XXXXXAAGCTGAATATTTTAATTTATGAC
gi|1730275|U34273          XXXXXAAGCTGAATATTTTAATTTATGAC
gi|1816508|U39634          ATGTAGAACTGAAAAAAAAAAAAAAAAAA
gi|1816508|U39634b         XXXXXAAGCTGAATATCTTAATTTATGAG
gi|1997|X68330             XXXXXAAGCTGACTATTTTAATTTATGAT
gi|3983428|AF104245        ATATAGACCTAAAAAAAAAAAAAAAAXX
gi|3983428|AF104245b       XXXXXAAGCTGAATATTTTAATTTATGAG
gi|4185290|AF083270        XXXXXAAGCTGAATATTTTAATTTATGAG
gi|52678|X05064            ACACGAATCTGAATGAGAATGCCTGTGAT
gi|52682|X05253            ACTTCATTGCCATAAGGTTCTACTGTTAG
gi|529238|L32955           ATAAAAAACAAACTTCCXXXXXXXXXXXX
gi|529238|L32955b          GTGTCCCACTGAAGGAGCAAGGCTCAGGC
```

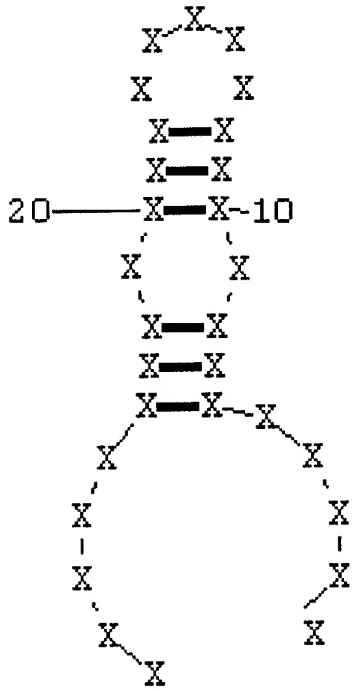
Figure 96

Score: 136.0



```
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN Consensus
UUAUAAUUAUUAUAAAAAUAAGUAUAUGU-     gi|1004228|X87408
XXXXXAAGCUGAAUAUCUUAAUUUAUGAG      gi|1004228|X87408b
AUGUAGAGCUGAAAAAAXXXXXXXXXXXX      gi|163210|M77120
XXXXXAAGCUGAAUAUUUUAAUUUAUGAG      gi|163210|M77120b
AAUUUAUGCUUUUUAAUAGCUUAUAU---      gi|165950|M96845
XXXXXAAGCUGAAUAUUUUAAUUUAUGAC      gi|165950|M96845b
XXXXXAAGCUGAAUAUUUUAAUUUAUGAC      gi|1730275|U34273
AUGUAGAACUGAAAAAAAAAAAAAAAAAA      gi|1816508|U39634
XXXXXAAGCUGAAUAUCUUAAUUUAUGAG      gi|1816508|U39634b
XXXXXAAGCUGACUAUUUUAAUUUAUGAU      gi|1997|X68330
AUAUAGACCUAAAAAAAAAAAAAAAAAXX      gi|3983428|AF104245
XXXXXAAGCUGAAUAUUUUAAUUUAUGAG      gi|3983428|AF104245b
XXXXXAAGCUGAAUAUUUUAAUUUAUGAG      gi|4185290|AF083270
ACACGAAUCUGAAUGAGAAUGCCUGUGAU      gi|52678|X05064
ACUUCAUUGCCAUAAGGUUCUACUGUUAG      gi|52682|X05253
AUAAAAAACAAACUUCCXXXXXXXXXXXX      gi|529238|L32955
GUGUCCCACUGAAGGAGCAAGGCUCAGGC      gi|529238|L32955b
```

Figure 97



DOCKET NO. IBIS-0012 **PATENT**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In Re Application of:

David J. Ecker, Ranga Sampath, Richard Griffey, John McNeil

Group Art Unit: Not Yet Assigned

Examiner: Not Yet Assigned

For: IDENTIFICATION OF MOLECULAR INTERACTION SITES IN RNA FOR NOVEL DRUG DISCOVERY

DECLARATION AND POWER OF ATTORNEY

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name; and

I believe that I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a

☒ Utility Patent ☐ Design Patent

is sought on the invention, whose title appears above, the specification of which:

☒ is attached hereto.

☐ was filed on ________________ as Serial No. ________________ .

☐ said application having been amended on ______________________ .

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose to the U.S. Patent and Trademark Office all information known to be material to the patentability of this application in accordance with 37 CFR § 1.56.

I hereby claim foreign priority benefits under 35 U.S.C. § 119(a-d) of any **foreign**

application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of any application on which priority is claimed:

Priority Claimed (If X'd)	Country	Serial Number	Date Filed
☐			
☐			
☐			
☐			

I hereby claim the benefit under 35 U.S.C. § 120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of 35 U.S.C. § 112, I acknowledge the duty to disclose to the U.S. Patent and Trademark Office all information known to be material to patentability as defined in 37 CFR § 1.56 which became available between the filing date of the prior application and the national or PCT international filing date of this application:

Serial Number	Date Filed	Patented/Pending/Abandoned
09/076,440	May 12, 1998	Pending

I hereby claim the benefit under 35 U.S.C. § 119(e) of any United States provisional application(s) listed below:

Serial Number	Date Filed
60/085,092	May 12, 1998

I hereby appoint the following persons as attorney(s) and/or agent(s) to prosecute this application and to transact all business in the Patent and Trademark Office connected therewith:

John W. Caldwell	Registration No. 28,937
Paul K. Legaard	Registration No. 38,534

of **WOODCOCK WASHBURN KURTZ MACKIEWICZ & NORRIS LLP,** One Liberty Place - 46th Floor, Philadelphia, Pennsylvania 19103 and

Herb Boswell	Registration No. 27,311
Laurel Bernstein	Registration No. 37,280
Henry Wu	Registration No. 44,412

of **ISIS PHARMACEUTICALS, INC.,** 2292 Faraday Avenue, Carlsbad, California 92008

Address all telephone calls and correspondence to:

Paul K. Legaard
WOODCOCK WASHBURN KURTZ
MACKIEWICZ & NORRIS LLP
One Liberty Place - 46th Floor
Philadelphia PA 19103
Telephone No.: **(215) 568-3100**
Facsimile No.: **(215) 568-3439**

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Name: David J. Ecker	
Mailing Address: 1041 Saxony Road Encinitas, California 92024	**Signature** **Date of Signature:** _____
City/State of Actual Residence: Encinitas, California 92024	**Citizenship:** United States

Name: Ranga Sampath	
Mailing Address: 12180 Ragweed Street San Diego, California 92129	**Signature** **Date of Signature:** _____
City/State of Actual Residence: San Diego, California 92129	**Citizenship:** United States

Name: Richard Griffey	
Mailing Address: 360 Barsby Street Vista, California 92084	**Signature** **Date of Signature:** _____
City/State of Actual Residence: Vista, California 92084	**Citizenship:** United States

Name: John McNeil	
Mailing Address: 427 Retaheim Way La Jolla, California 92037	______________________ **Signature** **Date of Signature:** ____________
City/State of Actual Residence: La Jolla, California 92037	**Citizenship:** United States

Name:	
Mailing Address:	______________________ **Signature** **Date of Signature:** ____________
City/State of Actual Residence:	**Citizenship:** ____________

Name:	
Mailing Address:	______________________ **Signature** **Date of Signature:** ____________
City/State of Actual Residence:	**Citizenship:** ____________